INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet symbiose

SYstèmes et Modèles BIOlogiques, BIOinformatique et SEquences

Rennes

THÈME 3A

Table des matières

1. Composition de l'équipe

Le projet Symbiose a été créé le 1er janvier 2002. Notre problématique générale correspond au champ de la bioinformatique, c'est à dire à la modélisation et à l'analyse de données génomiques et post-génomiques, afin d'assister le biologiste moléculaire dans la formulation et la découverte de nouvelles connaissances biologiques. Par définition, notre projet, centré sur un domaine applicatif, ne s'inscrit pas parfaitement dans la classification proposée par l'Inria. Le rattachement le plus naturel reste le thème « bases de connaissances » du programme 3A de l'Inria, en regard de l'objectif final d'améliorer l'accès à, d'expliciter et de découvrir des connaissances.

Responsable scientifique

Jacques Nicolas [CR Inria]

Assistante de projet

Maryse Auffray [AA Inria]

Personnel Inria

Rumen Andonov [Professeur, université de Valenciennes, en détachement Inria]
François Coste [CR Inria]
Michel Le Borgne [MC, université de Rennes 1, en détachement Inria]

Personnel CNRS

Dominique Lavenier [DR (à partir d'octobre 2002) CNRS]
Frédéric Raimbault [MC, université de Bretagne Sud, en délégation CNRS depuis septembre 02]
Anne Siegel [CR CNRS]

Personnel Université

Catherine Belleannée [MC, université de Rennes 1]
Israël-César Lerman [Professeur, université de Rennes 1]
Basavanneppa Tallur [MC, université de Rennes 1]
Raoul Vorc'h [MC, université de Rennes 1]

Chercheurs doctorants

Daniel Fredouille [allocataire MENRT, ATER depuis octobre 02]
Roberto Bonato [allocataire MENRT jusqu'à avril 02]
Stéphane Guyetant [allocataire BDI CNRS/Région]
Aurélien Leroux [allocataire Inria/Région]
Ingrid Jacquemin [allocataire MENRT]
Andre Floëter [cotutelle université Potsdam]
Yoann Mescam [allocataire Inria cofinancée SIB Genève]
Mathieu Giraud [normalien, depuis septembre 02]

Personnel sous contrat

Cynthia Alland [Ingénieur Associée, jusqu'à octobre 02]
Yves Bastide [Ingénieur Expert Inria (contrat européen StressGenes)]
Esther Kaboré [Ingénieur Expert Inria depuis septembre 02 (contrat Inria/Région Génopole)]
Hugues Leroy [Personnel Inria, à 30%]
Michel Mac Wing [Ingénieur Expert Inria (ACI GénoGrid)]
Emanuelle Morin [Ingénieur Expert Inria depuis septembre 02 (contrat Inria/national Génopole)]
Elodie Retout [Ingénieur CDD Inra (programme Inra Agenae)]

Visiteur

Nicolas Yanev [Professeur invité, université de Sofia (Bulgarie), accueil Inria, juin-septembre 20002]

Stagiaires

Raed Abdel-Malek [ESIB (Liban)]

Manuel Bes [DEA Génomique et Informatique]
Gaurav Chatley [IIT (Inde)]
Sophie Durand [DES (Université de Bruxelles)]
Thomas Henry [DEA Génomique et Informatique]
Armand Khoury [ESIB (Liban)]
Nathalie Leclerq [Maitrise d'informatique de Valenciennes]
Julien Pley [DEA IFSIC]
Solen Quiniou [3ème année INSA]
Pascaline Tchienhom [DEA IFSIC]
Philippe Veber [magistère MMMI]

2. Présentation et objectifs généraux

2.1. Un projet de bioinformatique

Les données que nous manipulons sont essentiellement des séquences, d'ADN, d'ARN ou de protéines, issues des banques de données publiques. Il s'agit de mots sur un alphabet à 4 ou 20 lettres auxquels sont rattachées des annotations parfois extrêmement complexes, décrivant les structures et fonctions associées au niveau des organismes, ainsi que de nombreux liens pointant sur les articles scientifiques associés ou reliant les objets d'étude entre eux . À ces données s'ajoutent les expérimentations des laboratoires de biologie qui portent pour la plupart sur ce que l'on dénomme maintenant la post-génomique. Contrairement à la génomique, qui considère l'ensemble des gènes d'une espèce, la post-génomique s'intéresse à l'expression de ces gènes dans des conditions et des tissus donnés pour un individu particulier, ceci tant au niveau de l'ARN que des protéines. Les données brutes sont alors des images que l'on convertit en niveaux d'expression ou en ensembles de masses. Elles sont donc de nature assez différente des précédentes, et impliquent des problèmes spécifiques pour leur traitement.

Le développement d'un tel type de projet n'est pas contradictoire avec une coloration bien spécifique des compétences que nous affichons par rapport à d'autres projets également axés sur la bioinformatique. En particulier, nous appuyons nos efforts sur la recherche et la découverte de signatures de famille de séquences en privilégiant la modélisation par langages formels et en ayant recours à des machines spécialisées pour ces traitements.

2.2. Thèmes scientifiques

Les *thèmes scientifiques* sur lesquels se focalise le projet découlent de notre choix de modéliser des systèmes biologiques complexes, en se plaçant dans un cadre linguistique et logique. De façon plus précise, le projet s'articule en trois grandes directions.

2.2.1. Analyse linguistique de séquences

Cet axe concerne la recherche des structures spatiales ou logiques pertinentes (i.e. fonctionnelles) dans les macro-molécules, qu'il s'agisse de modéliser des mécanismes biologiques généraux (transcription, épissage, frameshift...) ou des structures spatiales spécifiques (structures secondaires ou tertiaires de familles de protéines). Nous abordons ces problèmes dans le cadre de la théorie des langages, en nous posant des questions à la fois théoriques (comparaison de deux mots, classe utile de langages, classe apprenable) et pratiques (comment construire des analyseurs efficaces, comment inférer des langages à partir d'échantillons de phrases ?). Globalement, notre spécificité est le traitement de données de manière combinatoire, c'est-à-dire en s'appuyant sur le dénombrement de structures similaires pour rassembler ou caractériser plutôt que sur l'estimation ou l'adaptation de paramètres dans des modèles fixés. Les champs disciplinaires abordés sont les grammaires logiques, l'apprentissage automatique et l'analyse de données.

2.2.2. Analyse et identification de systèmes dynamiques

Le point précédent s'intéresse à la caractérisation pour chaque organisme de leurs gènes pris individuellement. Le but final reste néanmoins d'intégrer toutes ces données en un modèle de fonctionnement global, explicitant et permettant de simuler les interactions majeures entre gènes dans des environnements donnés. Cette démarche est indispensable à la pleine compréhension de pathologies ou de mécanismes de régulation. Du fait de la difficulté d'obtenir des données expérimentales précises et de la complexité des mécanismes étudiés, on ne peut espérer développer actuellement dans ce domaine que des modélisations de nature qualitative. Nous visons l'aide à l'identification de tels modèles. Notre approche consiste à utiliser une modélisation qualitative, à la valider au moyen de techniques issues de la vérification de circuits et plus généralement de l'automatique, puis à la raffiner par apprentissage automatique (identification de systèmes à évènements discrets). Les champs disciplinaires abordés sont l'analyse de données, l'apprentissage automatique et l'automatique des systèmes à évènements discrets.

2.2.3. Parallélisme

Les traitements que nous venons de décrire demandent une puissance de calcul d'autant plus grande que les données génomiques sont produites à un rythme extrêmement soutenu : doublement de la taille des banques publiques tous les ans, voire tous les 10 mois. L'accès rapide à des sélections complexes portant sur des millions d'objets, ou l'extraction de connaissances (*data mining*) sur ces mêmes objets devient alors un enjeu scientifique stratégique. L'objectif principal de cet axe de recherche est de paralléliser ces traitements pour en accélérer fortement l'exécution. La mise en oeuvre vise plusieurs catégories de machines : les super calculateurs, les grilles de calcul et les machines spécialisées. Les champs disciplinaires abordés sont le parallélisme, le *grid computing* et l'architecture des machines.

2.3. Des collaborations actives

2.3.1. La Génopole Ouest

La Génopole Ouest, créée en janvier 2002, implique les régions Bretagne et Pays de Loire, et coordonne les actions des différents laboratoires impliqués en génomique, post-génomique et bioinformatique dans le Grand Ouest. J. Nicolas est responsable du comité bioinformatique qui est en charge de la coordination de la gestion et du traitement des données de la Génopole, ainsi que des recherches effectuées en bioinformatique. Nous avons dans ce cadre développé des collaborations avec la plupart des biologistes concernés de l'université de Rennes, ainsi qu'avec le réseau de bioinformatique de la génopole, essentiellement à Nantes, Roscoff, Brest et Angers.

2.3.2. Aspect développement

Le projet présente un aspect applicatif important, nécessaire pour identifier puis formaliser les problèmes difficiles de biologie où l'approche « in silico » sera déterminante. Ceci suppose un travail d'ingénieur conséquent, pour le déchiffrement initial des problèmes et l'obtention de données biologiques pertinentes, puis pour l'application concrète des algorithmes développés en recherche. Nous fonctionnons pour cela avec des postes de CDD (obtenus dans le cadre de différents contrats). Hugues Leroy, Ingénieur Inria, collabore également avec le projet Symbiose pour les aspects grilles de calcul, à travers sa participation à l'ACI Genogrid.

2.3.3. Collaboration avec le projet TexMex

Afin d'obtenir des résultats pertinents en bioinformatique, on se heurte d'emblée au problème fondamental de l'acquisition de données de bonne qualité. En effet, bien des travaux échouent ou aboutissent à des résultats biologiquement peu significatifs du fait de l'extrême difficulté de recueillir des données propres dans le domaine. Nous collaborons avec le projet TexMex, qui a abordé sur ce thème deux directions de recherche. La première, provient de l'hétérogénéité, de la redondance et du manque de cohérence des diverses sources de données disponibles. Il s'agit alors de gérer et synchroniser des méta-données sur ces sources, puis d'établir la qualité des données rapatriées. La seconde difficulté découle de la nature même d'une des sources actuelles de données les plus riches, les articles associés aux séquences publiées. Il n'est plus envisageable de récupérer entièrement manuellement les informations publiées dans les nombreux journaux du domaine et l'analyse

et la compréhension de textes en langage naturel devient un passage obligé de l'acquisition de données. Ce problème sera abordé dans TexMex par analyse sémantique de résumés d'articles.

3. Fondements scientifiques

3.1. Bioinformatique

La bioinformatique est un terme aux contours suffisamment flous pour que nous précisions le cadre restreint dans lequel nous l'abordons. Les anglo-saxons distinguent en général deux termes : « biocomputing », qui désigne l'ensemble des travaux informatiques nécessaires pour gérer des données en masse au quotidien et « computational biology », plus spécifiquement utilisé pour la recherche à l'interface entre informatique et biologie, telle que nous la concevons. Notre expérience est qu'il est actuellement difficile d'effectuer des recherches en profondeur dans le domaine sans participer à l'activité de service correspondant au premier terme, du fait de la pauvreté des infrastructures informatiques en biologie.

Du point de vue de la biologie, les enjeux principaux de la bioinformatique sont l'aide à la découverte de cibles diagnostiques et thérapeutiques et la compréhension des mécanismes d'action du vivant. Ceci peut recouvrir en pratique de très nombreux travaux, et nous nous limitons ici à ce qui constitue le cœur de la discipline, c'est-à-dire l'analyse des trois macro-molécules biologiques fondamentales du vivant que sont l'ADN, l'ARN et les protéines. Le but est de comprendre la structure, la dynamique et les relations qui existent entre ces molécules, que ce soit au cours de l'évolution, dans un mécanisme biologique général ou dans un métabolisme particulier. On peut distinguer quatre grandes classes de travaux (pour de plus amples développements, on pourra se référer à la partie introductive de [55] ou à [71]) :

- *L'obtention des données.* Le champ des recherches semble restreint à ce niveau. Le problème principal demeure la reconstruction d'une séquence complète à partir de l'observation de fragments de séquences coupés au hasard dans le génome. Actuellement, les données ne sont plus simplement des textes caractérisant des séquences, mais des images haute définition représentant l'activité d'une cellule à un instant et dans un contexte donné. On est alors confronté à des problèmes d'analyse d'image pour ses données post-génomiques, qui semblent relativement classiques (détection de contours, segmentation, ...).
- *La gestion des données obtenues.* C'est actuellement un problème majeur. Les informations sont produites au niveau de chaque laboratoire avec une normalisation relativement lâche et les banques de données sont encore peu structurées, redondantes et incohérentes (on est généralement loin de véritables bases de données, sans parler de bases de connaissances). La méthodologie XML semble se développer fortement dans ce domaine. L'accès à l'information peut dans ces conditions devenir un véritable enjeu de recherches. En plus de la formalisation, le problème difficile est celui de l'intégration de toutes ces sources hétérogènes.
- *La comparaison des séquences.* Se référer à l'ensemble des séquences déjà connues est la première et principale méthode pour étudier de nouvelles séquences. Le problème fondamental est celui de l'alignement d'un groupe de séquences, où il s'agit de mettre en correspondance des positions jouant un même rôle structurel et/ou fonctionnel pour l'ensemble des séquences du groupe. Un des buts les plus importants dans ce contexte est l'établissement de phylogénies, c'est-à-dire d'arbres retraçant l'évolution des espèces. Les données de synthénie, c'est à dire d'agencement similaires de gènes à travers différentes espèces donnent une vue plus macroscopique de l'évolution des séquences, qui peut être utile également dans les stratégies de séquençage.
- *L'analyse fonctionnelle et structurelle des données génomiques.* Il s'agit d'un domaine vaste, en pleine expansion, qui vise à extraire des connaissances de la quantité des données récoltées sur les génomes, transcriptomes, protéomes, interactome et métabolomes de différents organismes. Il regroupe la recherche des gènes, la recherche de sites fonctionnels, de structures particulières et, de plus en plus, l'étude des interactions entre les différentes macro-molécules, en particulier dans les mécanismes de régulation.

Nos travaux se situent principalement dans le dernier axe présenté. Nous nous intéressons également à la comparaison de séquences dans les aspects calcul intensif qu'elle requiert.

3.1.1. Intérêt biologique de la découverte de motifs

Du fait de son importance dans le projet, nous détaillons un peu plus la motivation biologique du problème de la découverte des séquences. Les séquences biologiques, qu'il s'agisse d'ADN, d'ARN ou de protéines, doivent respecter un certain nombre de contraintes qui vont être extrêmement importantes pour la structure et/ou la fonction ou l'activité que doit exercer cette séquence in fine. Ces contraintes se traduisent par la conservation au cours de l'évolution de « motifs » plus ou moins précis et plus ou moins complexes [1]. La complexité peut varier de la présence de quelques lettres à une position donnée de la séquence à des relations de longue distance entre des mots, dues au repliement spatial des molécules, avec des phénomènes de symétrie, de copie, d'approximation... La conservation des motifs va permettre non seulement de caractériser une famille de séquences donnée, mais également d'expliquer dans une certaine mesure les rapports structure/fonction. Bien entendu, on ne peut être complètement à l'abri du repérage d'une zone qui fortuitement a peu muté et est restée conservée au cours de l'évolution. C'est pourquoi un retour à la paillasse et à l'expérimentation biologique reste toujours nécessaire pour valider la pertinence des motifs observés. Ces motifs, constitués à la main ou automatiquement, sont ensuite mis à disposition de la communauté dans des banques comme Prosite ou eMOTIF pour les protéines [2] ou TRRD pour l'ADN [3], où à travers des programmes de prédiction de sites biologiquement importants (transition intron/exon, cadre ouvert de lecture,...). Leur connaissance peut être utilisée dans de multiples applications en biologie et nous décrivons rapidement les principales. Un des tout premiers intérêts est la caractérisation de familles de protéines. Bon nombre de laboratoires se spécialisent en effet sur l'étude d'une famille particulière de protéines, intéressantes par leur structure, leur localisation, leur fonction ou leur implication dans un mécanisme pathologique. Travaillant sur quelques protéines, ils peuvent ensuite amplifier leurs découvertes en recherchant dans les banques publiques toutes les protéines répondant aux motifs trouvés. En ce qui concerne l'ADN, c'est plutôt les zones importantes de régulation, situées en amont des gènes, qui vont bénéficier d'un certain degré de conservation et la découverte de motifs dans ces zones fournira des renseignements importants à la fois sur la localisation probable des gènes et sur le degré d'expression de ces gènes. Un autre intérêt est de pouvoir réaliser des alignements multiples plus fiables sur les séquences (à condition que la méthode d'identification des motifs ne repose justement sur une méthode d'alignement multiple !), c'est à dire mettre en correspondance les acides d'une séquence à l'autre, ce qui permet ensuite de retracer plus facilement les transformations (mutations, réarrangements) les plus probables de ces séquences au cours de l'évolution. Enfin, ces motifs permettent l'annotation de protéines, c'est à dire de disposer d'indices pour prédire la fonction, l'activité ou la localisation d'une protéine nouvellement séquencée par simple observation de cette séquence. Ce travail est plus complexe que la caractérisation de familles, car les protéines présentent la plupart du temps plusieurs domaines (fréquemment trois ou plus) avec des motifs différents dont l'agencement conduit à la fonction spécifique.

3.2. Analyse linguistique de séquences

Mots clés : *apprentissage automatique*, *analyse de données*, *grammaires logiques*, *inférence grammaticale*, *recherche et découverte de motifs*, *comparaison de séquences*, *extraction d'informations*.

3.2.1. Langages formels et séquences biologiques

Du point de vue des séquences, considérées comme des mots sur un alphabet d'acides nucléiques ou d'acides aminés, l'ensemble de contraintes structurelles et fonctionnelles superposées conduit à la formation d'un véritable langage dont la connaissance permettrait de prédire les propriétés des séquences. La théorie des langages formalise les notions fondamentales sous-jacentes aux phénomènes étudiés (degré d'expressivité, complexité de l'analyse, automates associés, algèbre sur les langages). Encore très peu d'auteurs ont exploré ce paradigme. On peut l'étudier selon deux points de vue :

[1] nous employons parfois le terme « signature » pour spécifier que ces motifs peuvent avoir une complexité arbitraire et étendent la notion usuelle en biologie de motif consensus

[2] http://www.expasy.org/prosite, http://motif.stanford.edu/emotif

[3] http://dragon.bionet.nsc.ru/trrd

- Un point de vue fondamental, où il s'agit d'étudier la classe de langages formels la plus adaptée à la description des phénomènes naturels observés. Les splicing systems de Head [56], ou H-systems, reproduisant le phénomène du crossing over, représentent un des formalismes les plus féconds à cet égard. Des théoriciens des langages comme A. Salomaa et Gh. Paun [75] ont également exploré ce que deviennent les questions classiques (complexité, décidabilité, langages stables...) si l'on considère les opérations naturelles sur les séquences biologiques (inversion, transposition, duplication, délétion,...) et mis au point en particulier un modèle appelé Sticker-system basé sur l'opération de complémentarité telle qu'on l'observe dans les appariements Watson Crick [64]. Leur objectif est de mettre au point des systèmes de calcul ayant la puissance des Machines de Turing, dans la ligne des travaux sur le DNA-computing, ce qui est un peu différent de la question de la classe des langages nécessaire pour décrire les structures biologiques réelles. On pense actuellement que l'expressivité nécessaire se situe dans la classe des langages « mildly context sensitive », bien connue des analyseurs de langage naturel. Par exemple Y. Kobayashi et T. Yokomori à Tokyo ont modélisé et prédit les structures secondaires d'ARNs à l'aide de grammaires d'arbre adjoints (TAGs) [92]. Les travaux les plus complets dans ce domaine nous semblent être ceux de D. Searls, de Smith KLine Beecham en Pennsylvanie[83][84] ;
- Un point de vue plus pratique, où il s'agit de fournir au biologiste le moyen de formaliser son modèle à l'aide d'une grammaire, grammaire qui, soumise à un analyseur, permettra ensuite d'extraire des banques de données publiques les séquences pertinentes vis à vis du modèle. J. Collado Vides a été un des premiers à s'intéresser à cette voie pour l'étude de la régulation des gènes [46]. D. Searls a proposé une approche plus systématique basée sur les grammaires logiques et un analyseur, Genlang [50], encore très peu utilisé dans la communauté des biologistes car exigeant des compétences avancées en langages. C'est cette solution à partir de laquelle nous avons démarré nos propres travaux, en mettant l'accent sur l'accessibilité du modèle aux biologistes, c'est à dire la possibilité pour ceux-ci de construire eux-mêmes les modèles et interpréter les résultats.

S'il est bon de connaître l'expressivité nécessaire pour décrire les modèles biologiques, le biologiste reste très souvent incapable de fournir de tels modèles. L'assister dans le travail de construction des modèles suppose de développer des techniques d'apprentissage automatique.

3.2.2. Découverte de motifs

Du fait de son importance pratique et de la quantité croissante de données disponibles, on a vu se multiplier ces dernières années les programmes d'apprentissage de motifs. On trouvera des revues du domaine dans [40] ou [59]. Le premier critère pour classer les méthodes est celui du type de motifs recherché et son expressivité. On peut essentiellement représenter un langage soit dans un cadre probabiliste, par une distribution sur l'ensemble des mots que l'on peut former sur le vocabulaire, soit dans un cadre de langages formels, par un système de génération de l'ensemble des mots acceptés. À l'interface entre les deux, on trouve les réseaux de Markov cachés et les automates stochastiques, qui ont de très bonnes performances, mais où classiquement la structure est figée et l'apprentissage porte sur les paramètres de la distribution. Ceci les rapproche donc plutôt du point de vue apprentissage du premier type de représentation. La représentation distributionnelle s'exprime selon différentes modalités : matrices consensus (probabilité d'occurrence à chaque position de chaque lettre), profils (prise en compte des gaps), matrices de poids (quantité d'information apportée par chaque position et contribution de chaque lettre). Au niveau algorithmique, la notion d'alignement joue en général un rôle fondamental. Des mots courts sont recherchés, puis des alignements effectués par programmation dynamique autour de ces points d'ancrage. La production de « blocs » est typique de cette approche [57]. Une recherche simplifiée des motifs peut alors être menée après alignement, les intervalles variables entre sous-motifs ayant été décidés. Les programmes les plus performants dans ce domaine semblent être actuellement Gibbs Motif Sampler du Wadsworth Center à New York, procédure bayésienne construisant une matrice consensus par échantillonnage de Gibbs [70] et Meta-MEME, à Columbia University, construisant un réseau de Markov de combinaison de telles matrices, produites par un algorithme EM (Expectation-Maximization).

La représentation linguistique, qui correspond à notre propre classe de travaux, repose généralement sur des expressions rationnelles. Les algorithmes utilisés sont alors de type énumération combinatoire dans un espace partiellement ordonné. Parmi les plus aboutis dans ce domaine, on trouve le programme Pratt, de l'université de Bergen [39], dont les principes sont très voisins de ceux établis par M.F. Sagot à Pasteur et A. Viari [80], ainsi que des variations sur la recherche de cliques dans un graphe [67][42].

Même si les résultats obtenus sont intéressants dans un certain nombre de cas, il y a selon nous une limitation fondamentale aux travaux actuels, qui restent tous assez fortement dépendants de la notion de position. C'est essentiellement la présence à une position donnée d'un certain type de lettre qui va conduire à la prédiction. Or il est clair que les relations existant entre divers sites, parfois éloignés sur la séquence, jouent un rôle biologique important. Prendre en compte ces contraintes structurelles passe par l'utilisation de modèles plus complexes. L'apprentissage purement statistique nous semble atteindre ses limites ici, du fait de la multiplication des paramètres à ajuster. Le cadre théorique qui nous semble le plus adapté à cette fin est celui des langages formels, où l'on peut chercher à optimiser cette fois comme paramètre principal la complexité de la représentation (principe de parcimonie). C'est cette voie dans laquelle nous sommes engagés, où découvrir un motif équivaut à apprendre un langage.

3.2.3. Apprentissage automatique et inférence grammaticale

Les travaux sur l'apprentissage ont connu à l'origine des développements dans deux communautés fort différentes : en psychologie expérimentale, où il s'agissait de comprendre et de simuler les mécanismes naturels d'apprentissage et en statistiques, où le but était de régler automatiquement les valeurs de paramètres de modèles en fonction de données observées. Optant pour une voie médiane algorithmique, une communauté d'apprentissage automatique s'est fortement développée durant les années 80 et occupe depuis lors une part croissante des travaux en IA. Elle possède deux grandes directions, une direction purement théorique (COmputational Learning Theory), où il s'agit d'étudier les critères d'apprenabilité et les classes de fonctions apprenables suivant ces critères, et une direction algorithmique (Machine Learning) où il s'agit de développer des algorithmes et les tester sur des données réelles ou simulées. On peut observer deux grandes tendances dans ce domaine. La première est une *volonté de rapprochement entre les communautés théoriques et expérimentales*, avec notamment des résultats sur les techniques de boosting (comment améliorer automatiquement un algorithme de prédiction ayant un taux de prédiction initial meilleur que celui d'un algorithme aléatoire) et l'apprentissage par vecteurs supports (transformation de l'espace de représentation pour obtenir une meilleure séparabilité des classes en discrimination). La seconde tendance concerne un *recours croissant à des techniques d'origine statistique* (apprentissage par renforcement, classification, physique statistique), auxquelles on peut adjoindre les travaux utilisant des réseaux connexionnistes ou des réseaux de Markov cachés (HMMs). Le problème de la comparaison et de l'intégration de méthodes dites symboliques-numériques a ainsi fait l'objet de nombreux travaux [6].

Actuellement, des approches comme les HMMs sont très utilisées et remportent des succès pratiques certains. L'aspect algorithmique y est relativement figé, la programmation dynamique représentant la plupart du temps le cœur des méthodes. Les raffinements portent sur la conception de mesures de vraisemblance ou de coefficients d'association, car on cherche avant tout à caractériser une distribution. Autrement dit, on s'intéresse plus à l'obtention d'un modèle statistiquement fidèle, qu'à la caractérisation d'un modèle explicable, fournissant une abstraction réaliste du système. Le problème dur, et pour lequel les applications sont potentiellement nombreuses, reste celui de la découverte de relations ou de structures dans un ensemble de données et c'est pourquoi nous privilégions ce domaine. Traditionnellement abordé par l'analyse des données, il réclame des idées nouvelles lorsque les données sont difficilement réductibles à des vecteurs attributs-valeurs. En particulier, se pose le problème de l'élaboration de mesures de similarité et de coefficients d'association, ainsi que celui du traitement de la complexité engendrée par le système descriptif.

Dans le domaine de la bioinformatique qui nous intéresse, les séquences constituent ainsi des données importantes et difficiles à traiter. Un champ de l'apprentissage s'intéresse particulièrement à ce type de données. Plus précisément, on appelle inférence grammaticale l'apprentissage automatique d'un modèle de langage à partir d'un échantillon fini des phrases du langage qu'elle accepte (instances positives) et

éventuellement d'un échantillon fini de phrases n'appartenant pas à ce langage (instances négatives). Notons qu'interviennent à la fois une notion de classe de langage, et une notion de représentation de ces langages.

Spécifier complètement un problème d'inférence grammaticale suppose de

- définir un alphabet pertinent et une classe des langages acceptés ;
- choisir une représentation des langages de la classe puis définir une relation d'ordre (relation de généralité) sur ces représentations, compatible avec l'inclusion sur les langages ;
- enfin, spécifier une stratégie d'exploration de l'espace des représentations choisi, fondée sur l'utilisation de l'ordre, les propriétés de la classe et éventuellement du vocabulaire utilisé et des connaissances à priori disponibles.

L'étude de l'inférence grammaticale correspond à une communauté relativement restreinte, avec des écoles espagnole (université polytechnique de Valence, université d'Alicante) et japonaise (Fujitsu labs) particulièrement actives. En France, quatre autres équipes travaillent dans le domaine, avec lesquelles nous avons des échanges réguliers, à St Étienne, Lille et Marseille (aspects théoriques) et à Lannion (application à la parole).

L'état de l'art correspond à l'apprentissage d'un langage régulier à partir d'exemples et contre-exemples de mots appartenant à ce langage [81]. Il existe également de nombreuses études sur des sous-classes de grammaires algébriques. Nos travaux visent plutôt à faciliter l'utilisation pratique des algorithmes d'inférence de langages réguliers, en relation avec les contraintes soulevées par les problèmes de bioinformatique : inférence multi-langages, acceptation du non déterminisme, vocabulaire structuré, recherche de solutions multiples, extension des capacités d'apprentissage par prétraitement des données et post-traitement des automates inférés...

3.3. Modélisation, analyse et simulation de systèmes dynamiques

La modélisation des interactions au niveau cellulaire est déjà ancienne en biologie. Elle a d'abord été l'oeuvre de biochimistes souhaitant connaître la dynamique de systèmes enzymatiques. Nous pouvons citer dans ce domaine un modèle très évolué du globule rouge paru récemment [60]. Ce modèle est l'aboutissement de plus d'une décennie de travaux en biochimie. Le globule rouge étant dépourvu de noyau, le modèle est essentiellement basé sur la cinétique des réactions enzymatiques.

Du côté de la génétique, dès que la présence de mécanismes de régulation des gènes a été démontrée, sont apparus des modèles simples des interactions géniques.

Le degré le plus grossier de la modélisation est la vision d'un réseau de régulation de gènes comme un graphe orienté dont les arcs étiquetés $+$ ou $-$ indiquent des activations ou des inhibitions. Les bases de données sur les interactions de régulation comme KEGG [63], GeneNet[68], EcoCyc[65], RegulonDB[82], qu'elles soient généralistes comme les deux premières ou plus spécialisées comme les deux dernières, contiennent, au moins implicitement, une description sous forme de graphes des interactions connues.

Notons également que la représentation sous forme de graphe est à la base du formalisme des réseaux bayésiens. Ce formalisme a été utilisé récemment avec un certain succès pour l'identification (au sens apprentissage) de réseaux d'interactions à partir de données expérimentales provenant de micro-arrays[51]. Un des avantages de cette technique est de reposer sur des bases statistiques solides. Elle présente des limitations sur les aspects dynamiques des interactions liées à la pauvreté des données expérimentales (absence de séries temporelles de taille significative).

Dans l'échelle croissante de la complexité, et parmi les modèles les plus anciens, on trouve ceux qui s'inspirent des circuits électroniques [66]. Chaque gène y est modélisé par une variable booléenne, qui est une fonction de l'ensemble ou d'un sous-ensemble des variables. La modélisation des réseaux d'interaction géniques par des réseaux booléens de Kaufman a deux faiblesses : le caractère binaire de l'état de chaque gène et le synchronisme (tous les gènes changent d'état au même moment). L'utilisation de logiques multivaluées ainsi que l'introduction d'automates indéterministes permet de s'affranchir de ces contraintes. Ces nouveaux modèles logiques peuvent également être vus comme des approximations qualitatives des modèles basés

sur des équations différentielles linéaires par morceaux [90][89][73]. Il est possible de montrer, avec des hypothèses un peu restrictives, que les propriétés du modèle logique sont en correspondance avec les propriétés des systèmes différentiels (états stationnaires, cycles limites...) [85]. Ces systèmes booléens généralisés ont connu un certain succès dans l'analyse de réseaux de régulation génique de faible taille (<20 gènes) ainsi que leur simulation[49].

Le développement des techniques de vérification de circuits électroniques a permis de manipuler des fonctions logiques de plusieurs centaines de variables[41][45]. La grande similitude avec les modèles booléens utilisés pour modéliser les réseaux d'interactions géniques laisse penser que leur utilisation est possible dans ce domaine avec comme bénéfice un saut dans la taille des réseaux analysables. Un des problèmes ouverts est la pertinence biologique des propriétés qui sont exprimables en logique temporelle par exemple, propriétés dont la vérification est devenue routinière dans le cadre de la vérification de circuits[47].

Bien entendu, les modèles à base d'équations différentielles ordinaires ont été étendus aux réseaux de régulation géniques. On obtient ainsi des modèles expliquant le cycle circadien ainsi que de nombreux phénomènes périodiques, des modèles du cycle de division des cellules ainsi que de bien d'autres phénomènes biologiques. La théorie des systèmes dynamiques (variété stable/instable, analyse des bifurcations...) est l'outil mathématique de base dans ce domaine.

Bien d'autres types de modèles ont été utilisés pour formaliser les réseaux d'interactions géniques : équations différentielles stochastiques, réseaux de Petri hybrides, équations différentielles qualitatives, systèmes à base de règles.... Notons récemment une proposition d'utiliser une variante du pi-calcul de Milner pour modéliser notamment l'aspect localisé des réactions dans la cellule ainsi que la mobilité des composants interagissant[79]. Le fait que deux produits ne peuvent pas entrer en réaction s'ils ne sont pas dans le même compartiment cellulaire, et les phénomènes de transport et délais qui en découlent, sont rarement pris en compte dans les modèles. Il est vrai que les modèles qui ne s'intéressent qu'aux interactions entre gènes par l'intermédiaire de leurs produits, ignorent totalement toutes les phases intermédiaires qui peuvent être très complexes. Abstraire ces étapes intermédiaires est un des problèmes qui doit être résolu pour exploiter les connaissances acquises en biochimie et physiologie.

Une revue des travaux sur la modélisation et la simulation de systèmes de régulation au niveau génétique est parue[48]. On y trouvera des descriptions plus détaillées des différents modèles. Par contre les techniques développées pour les modèles booléens dans le cadre de la vérification de circuits sont ignorées.

3.4. Parallélisme

Mots clés : *architectures parallèles*, *grilles de calcul*, *architectures spécialisées*, *architectures reconfigurables*.

L'usage du parallélisme en génomique trouve à la fois ses motivations dans le volume des données à traiter et dans la complexité des algorithmes mis en jeu. À la base, il y a les données issues du séquencage des génomes. À ce jour (fin 2002), plus de 110 génomes (beaucoup plus si on considère toutes les bactéries plasmides et organelles séquencées) - dont celui de l'homme - ont été entièrement séquencés. À cela, il faut rajouter les données en cours de séquencage de nombreux autres organismes, plus de 500 à l'heure actuelle selon la *Genomes onlines database*[4] . Ces données sont emmagasinées dans des banques dont le volume double tous les ans, voire tous les 9-10 mois. La progression est donc exponentielle et rien ne semble indiquer un fléchissement pour les années qui viennent.

Se pose alors le problème de la recherche d'informations dans ces banques, notamment l'interrogation à partir de séquences inconnues pour en extraire des éléments similaires. En fait, c'est une des tâches de routine de la génomique que de mettre en évidence des similarité entre séquences pour aider à la compréhension des fonctionnalités des macro-molécules que sont les protéines. L'hypothèse de base est que deux séquences (ADN ou protéines) qui se ressemblent ont de forte chance d'appartenir à une même famille, donc d'avoir des fonctionnalités équivalentes ou proches.

[4] http://ergo.integratedgenomics.com/GOLD

Les premiers algorithmes de comparaison de séquences sur la base de techniques de programmation dynamique ont été développés dans les années 70 [74]. Par la suite, diverses heuristiques ont été apportées pour l'implémentation de méthodes plus rapides comme BLAST [87]. Si ces dernières apportent un gain très appréciable sur les temps de calcul, il en résulte une approximation des résultats qui, dans certains contextes, sont préjudiciables. Une incitation forte à paralléliser les traitements les plus précis - mais aussi les plus coûteux en temps de calcul - s'est donc fait rapidement sentir. On peut citer les travaux de JJ. Codani, à l'Inria, qui, à travers le logiciel LASSAP [53], intègre un jeu de logiciels standards et reconnus d'alignement de séquences sur des calculateurs parallèles.

Dans les mêmes temps, d'autres voies de recherche, notamment la définition de structures matérielles spécialement adaptées à ce type de traitement ont également été explorées. Plusieurs prototypes d'accélérateurs matériels tels que SAMBA [7], BISP [44] ou BioScan [91], ont été imaginés et ont conduit à des produits commerciaux performants comme par exemple les accélérateurs BioXL, DECYPHER et GeneMatcher des sociétés Compugen ltd.[5], TimeLogic[6] et Paracel[7]. Ces machines peuvent comprendre des centaines, voire des milliers de processeurs. Elles sont dédiées à la recherche rapide de similarité dans les banques de séquences génomiques (ADN ou protéines).

Au delà de la simple recherche dans les banques, cette grande quantité d'information mise à disposition des chercheurs ouvre naturellement d'autres pistes d'investigation encore plus consommatrices de temps de calcul comme, par exemple, la comparaison de génomes entiers, la classification de toutes les protéines connues (projet décrypton), la constitution de bases de données spécialisées, telle ProDom, etc. Les solutions présentées précédemment pour accélérer les traitement restent plus que jamais d'actualité, même si depuis 2 ou 3 ans d'autres alternatives se profilent avec les grilles de calcul. Il s'agit de faire coopérer - pour un même traitement - un ensemble de calculateurs répartis géographiquement et connectés par Internet. Plusieurs projets de grille intégrant des applications de biologie, sont en cours parmi lesquels le projet DataGRID et son groupe de travail (WP10) bioinformatique, le projet BioGRID (une des quatre applications du projet EuroGRID), et le projet GénoGRID émanant d'une action concertée incitative nationale sur les grilles (ACI GRID) mise en place par le ministère en 2001.

Le volume des données génomiques n'est pas le seul facteur qui pousse à paralléliser les traitements. La complexité des algorithmes manipulant ces données en est un autre, en particulier les algorithmes impliqués dans la prédiction des structures tridimensionnelles des protéines. En effet, retrouver la conformation dans l'espace d'une protéine à partir de sa simple séquence d'acides aminés reste un défi sur lequel bon nombre d'équipes de recherche concentrent leurs efforts. Le défi se situe à la fois en terme de modélisation et en terme de résolution du problème en un temps raisonnable. À preuve, les moyens mis par IBM dans le projet *Blue Gene* (100 Million de dollars sur 5 ans) dans la réalisation d'un calculateur parallèle spécialisé pour ce type de problème[8]. Ce problème général se décline en plusieurs activités qui vont de la prédiction *de novo* (*protein folding*) à la reconnaissance de repliements (*protein threading*), en passant par la recherche de motifs 3D. La première méthode essaie de prédire la structure 3D de n'importe quelle protéine, la seconde tente de faire correspondre - si possible - une séquence protéique inconnue à une structure 3D connue, et la troisième cherche à extraire des motifs tridimensionnels communs. Les algorithmes qui résolvent ces problèmes sont d'une grande complexité (NP-complet). La parallélisation de ces algorithmes, que ce soit sur des supercalculateurs ou des grilles, est, de par leur importance, une activité en plein essor.

4. Domaines d'application

Mots clés : *biologie*.

Le projet Symbiose est centré sur le domaine de la bioinformatique. Les enjeux principaux sont l'aide à la découverte de cibles diagnostiques et thérapeutiques et la compréhension des mécanismes du vivant.

[5] http://www.compugen.co.il/
[6] http://www.timelogic.com
[7] http://www.paracel.com
[8] http://www.research.ibm.com/bluegene

Le projet est bâti autour d'un constat : la compréhension des mécanismes du vivant à partir des génomes, transcriptomes et protéomes nécessite une chaîne de traitements complexe qui suppose la réunion de compétences biologiques et informatiques. Elle suppose également l'intervention de disciplines informatiques variées (analyse de données, base de données, apprentissage automatique, architectures parallèles, automatique). Notre conviction est qu'une équipe ne présentera des avancées significatives dans le domaine que si elle consent à un investissement relativement lourd. Un tel investissement correspond au temps de développement d'une culture commune entre biologistes et informaticiens, mais également à la mise en place d'une dynamique informatique entre thèmes traditionnellement peu connectés, unis pour la résolution d'un même ensemble de problèmes biologiques.

Dans ce but, nous avons réuni un ensemble de personnes issues d'horizons scientifiques différents. Nous avons également participé à la création de la Génopole Ouest, ce qui fournit un environnement particulièrement favorable au projet et lui permet un très bon ancrage par rapport à la communauté biologique.

Nous développons :

- un projet multidisciplinaire de bioinformatique, capable de s'attaquer à l'ensemble de la chaîne d'analyse des données de séquences génomiques et d'expression post-génomique ;
- des collaborations actives avec des laboratoires de biologie et une animation du thème à travers la participation au pilotage de la Génopole Ouest, l'organisation de séminaires et l'enseignement, ainsi que le développement d'une plate-forme d'outils bioinformatiques rendant notre savoir-faire accessible aux laboratoires de biologie.

Nos compétences concernent plus particulièrement la recherche de signatures complexes de familles de séquences qui permet aussi bien d'étudier les mécanismes de régulation des gènes que les sites actifs dans les protéines produits de ces gènes. Un objectif important est de pouvoir découvrir dans un génome le code de protéines de familles d'intérêt dont on connaît uniquement quelques exemplaires.

Nous sommes également intéressés par la recherche de structures complexes dans l'ADN ou l'ARN. Nous avons ainsi travaillé sur la modélisation du décalage du cadre de lecture lors de la traduction ou sur la caractérisation de transposons à l'intérieur de génomes de plantes.

Nous nous attaquons plus généralement aux applications nécessitant une approche combinatoire exigeante en calcul, pour lesquelles des architectures parallèles de calcul sont nécessaires. Entrent dans cette catégorie la recherche des structures spatiales dans les protéines (« protein threadings », recherche de ponts disulfurés ...) et la comparaison intensive au niveau de génomes entiers (que ce soit dans un but de recherche de séquences homologues ou pour rechercher un jeu complet d'amorces pour découper le génome ...).

Enfin, nous démarrons des recherches sur la modélisation de réseaux de régulation de gènes/protéines, dans le but d'en extraire des propriétés importantes pour la compréhension, voir le contrôle, de cette régulation.

5. Logiciels

5.1.1. Plate-forme d'extraction de motifs

Participants : Cynthia Alland, Emmanuelle Morin, Jacques Nicolas, Yoann Mescam.

L'amélioration des techniques de biologie moléculaire a permis d'accroître de manière importante ces dernières années le nombre de séquences biologiques disponibles. Ces séquences peuvent être soumises à diverses analyses et notamment à l'extraction de motifs qui vise à caractériser des ensembles de séquences appartenant à une même famille. Un inventaire des algorithmes déjà existant a été réalisé en les classant selon plusieurs dimensions, incluant en particulier la complexité du langage d'expression du motif.Des algorithmes ont été ou sont actuellement développés au sein de l'équipe. Nous avons ainsi pu sélectionner un panel d'algorithmes dont l'expressivité des motifs trouvés couvre un large spectre. Le but est de rendre l'ensemble de ces algorithmes accessible aux informaticiens et aux biologistes. Une plate-forme Web de recherche de motifs a été réalisée. Cinq algorithmes sont disponibles. Pour certaines méthodes nous avons essayé de réduire le

nombre de paramètres pertinents à régler et de mettre au point des filtres visant à réduire le nombre de résultats produit. Afin de faciliter l'interprétation des résultats, nous avons intégré des modules de visualisation, de recherche de motifs dans des banques de séquences biologiques, des génomes ou des séquences courtes.

L'amélioration des techniques de biologie moléculaire a permis d'accroître de manière importante ces dernières années le nombre de séquences biologiques disponibles. Ces séquences, utilisant un alphabet de 4 ou 20 lettres, contiennent des régions (dites motifs) plus importantes que d'autres du point de vue de la fonction biologique et que l'on retrouve conservées dans un groupe de séquences ayant une fonction commune. L'identification de ces motifs est fondamentale pour une meilleure compréhension des mécanismes cellulaires du vivant. Elle est aussi utile dans son aspect « boîte noire », à savoir la prédiction d'une fonction à partir de séquence inconnue. L'identification de ces motifs peut être assimilée à la recherche d'un langage commun pour l'ensemble des séquences.

C. Alland a réalisé une plate-forme Web regroupant des algorithmes d'extraction de motifs. Cette plate-forme sera utilisée par les biologistes pour une recherche plus fiable et plus rapide des motifs en comparant et associant les résultats de l'ensemble des méthodes disponibles mais aussi par les informaticiens pour comparer sur une base objective les performances en temps et en qualité de leurs algorithmes. Pour rendre cette plate-forme plus fonctionnelle, nous avons été amenés à développer d'autres modules pour l'exploitation et l'interprétation des résultats.

La plate-forme Web[9] est hébergée par un serveur Sun multiprocesseur au PCIO, Pole de Calcul Intensif Ouest. Les langages de programmation choisis sont Python, PHP et JavaScript. La disponibilité de la plate-forme, prévue en priorité pour les laboratoires de la Génopole Ouest, devrait être accessible par la suite de façon publique.

Un inventaire[40] des algorithmes déjà existant a été réalisé en les classant selon plusieurs dimensions, incluant en particulier la complexité du langage d'expression du motif variant de la classe A à I, I étant la plus générale. Nous avons sélectionné 8 algorithmes qui couvrent le plus largement possible le panel d'expression.

Afin de permettre aux biologistes d'interpréter les motifs trouvés par les algorithmes et d'affiner leur recherche, nous avons intégré à la plate-forme :

- un module de recherche de motifs dans les banques publiques accessible via les pages de résultats,
- un module *analyse des fragments inter-motifs*.

Plus précisément, les algorithmes implémentés sont :

- Algorithme de Staden[86].
- PRATT2[61][62].
- meta-Pratt, où nous proposons d'utiliser Pratt non plus sur des séquences d'acides aminés mais sur des séquences de motifs.
- Algorithme de Landraud[69].
- Smile[72].
- Winnower[76].

Nous avons aussi intégré un outil de recherche de motifs dans des grandes séquences développé dans l'équipe par Y.Mescam. Cet outil permet la recherche de motifs de classe F, avec la possibilité d'y intégrer des variables de chaîne. Actuellement, la recherche de motifs peut se faire sur les trois génomes complets qui sont stockés sur le serveur : l'homme, la souris et la drosophile.

La plate-forme est fonctionnelle et l'existence de modules distincts permet une intégration facile de nouveaux algorithmes. Lors de collaborations avec des biologistes nous avons pu utiliser la plate-forme sur trois exemples biologiques réels et avons pu illustrer l'utilité du couplage « extraction de motifs - recherche de motifs », en mettant en évidence des séquences pouvant coder pour de nouveaux membres de la famille étudiée. Le rapport concernant ce contrat est disponible sur le Web [30].

[9] http://idefix.univ-rennes1.fr:8080/PatternDiscovery/ Pour toute utilisation, contacter E. Morin

6. Résultats nouveaux

6.1. Analyse linguistique de séquences

Participants : Catherine Belleannée, François Coste, Jacques Nicolas, Raoul Vorc'h, Cynthia Alland, Emmanuelle Morin, Daniel Fredouille, Roberto Bonato, Aurélien Leroux, Ingrid Jacquemin, Yoann Mescam.

Deux types de travaux sont menés dans ce cadre. Lorsque le biologiste dispose d'un modèle, il s'agit de le rendre opérationnel pour le valider sur de très grandes bases de données. Nous nous fondons pour cela sur une analyse par grammaire logique. Lorsque aucun modèle n'est connu, nous essayons de l'inférer. Nous proposons des avancées théoriques et pratiques dans le domaine de l'inférence grammaticale, avec le but de démontrer sur quelques problèmes biologiques la faisabilité de la reconnaissance et de la découverte de signatures pertinentes complexes. Nous visons également la mise au point d'une plate-forme intégrant, outre ceux que nous développons, une palette la plus complète possible d'outils de recherche et de découverte de motifs complexes, facilement utilisables, en relation avec les principales banques disponibles.

6.1.1. Analyse par grammaires logiques

Participants : Catherine Belleannée, Jacques Nicolas, Cynthia Alland, Yoann Mescam.

Du fait qu'il a été le seul à aborder les aspects théoriques et pratiques de l'analyse de séquences, nous privilégions les travaux de Searls pour le développement de notre propre approche. Ceux-ci sont basés sur le formalisme des grammaires logiques et étendent les DCGs avec la notion de variable de chaîne et de morphisme.

Un des premiers enjeux est de rendre le formalisme accessible au biologiste, afin qu'il puisse lui-même, avec un minimum de formation, concevoir et tester ses modèles. Ceci suppose l'intégration d'un ensemble de connaissances biologiques « de sens commun » dans les grammaires, ainsi que la conception d'interfaces de programmation visuelle. La mise au point de modèles pour l'ADN, l'ARN ou les protéines devient alors assez différente du fait du type de structure manipulée, même si l'analyseur sous-jacent reste le même.

La difficulté est ensuite de proposer un compromis expressivité/complexité qui permette la mise au point d'analyseurs efficaces, et ceci plus particulièrement pour le traitement de génomes ou de chromosomes complets. Nous cherchons pour cela à exploiter au maximum la puissance d'une analyse lexicale capable de traiter de très grandes séquences. L'idée est de se fonder sur une structure de données de type arbre des suffixes, offrant des possibilités de calcul particulièrement souples. Un stage de DESS (S. Durand) a abouti à un premier outil d'analyse, capable de traiter les expressions Prosite et des répétitions élémentaires. Une expérimentation a été menée sur le génome d'Arabidopsis Thaliana pour y rechercher de manière systématique une famille de transposons [16].

6.1.2. Plate-forme de découverte de motifs

Participants : Jacques Nicolas, Cynthia Alland, Esther Kaboré, Emmanuelle Morin.

Notre ambition est de proposer une plate-forme intégrée de découverte de motifs, c'est à dire offrant à la fois une riche palette d'algorithmes de découverte pour des degrés d'expressivité variés dans un cadre unifié, et offrant un environnement permettant de sélectionner et tester effectivement la validité des motifs trouvés. Les aspects de développement sont bien entendu importants sur cette action, mais nécessaires pour proposer un véritable outil d'aide à la découverte, utilisable et utilisé. C. Alland, ingénieur associée Inria, se consacre actuellement à cette tâche. Ses travaux sont repris par E. Morin. Les travaux de recherche associés portent essentiellement sur les outils de filtrage de résultats. En effet, la plupart du temps, les algorithmes sont très prolifiques dans leurs résultats bruts et de ce fait peu exploitables en pratique, car le biologiste a besoin d'une vision synthétique des résultats. Ceci suppose un travail de post-traitement, qui passe par des tests de subsumption et de la classification. Des expérimentations ont été menées sur le génome humain, pour la reconnaissance et la recherche de la signature d'une famille particulière de protéines, en collaboration avec le laboratoire Inserm Germ à Rennes. De nouvelles protéines importantes pour leurs implications dans le système

de défense immunitaire ont été découvertes et sont en cours de validation au Germ (publication en cours de rédaction).

6.1.3. Inférence grammaticale : travaux théoriques

Participants : François Coste, Jacques Nicolas, Roberto Bonato, Daniel Fredouille, Aurélien Leroux.

Nous avons provisoirement clos en début d'année un travail sur l'inférence de transducteurs, suite à l'arrêt de la thèse de R. Bonato.

Nous avons poursuivi l'étude de l'utilisation d'automates non déterministes pour l'apprentissage de caractérisations de séquences biologiques. Les principaux résultats obtenus affinent la caractérisation de l'espace de recherche : nous avons exhibé un ordre lui conférant une structure de treillis et développé l'étude de deux opérateurs de parcours associés, d'abord la fusion pour déterminisation (classiquement utilisée dans certains algorithmes d'apprentissage d'automates) mais aussi un nouvel opérateur plus particulièrement adapté au cas non-déterministe, la fusion pour désambiguisation, qui permet l'apprentissage d'automates non ambigus. Ces résultats permettent d'envisager de nouveaux modes de parcours de l'espace de recherche utilisant les résultats existant sur l'exploration et la représentation des treillis et ainsi d'envisager de nouveaux types d'algorithmes d'apprentissage dans les cas déterministes ou non.

En collaboration avec N. Yanev, nous avons étudié l'utilisation des techniques de programmation linéaire en nombres entiers pour attaquer le problème du plus petit automate déterministe compatible avec des exemples positifs et négatifs. Nous avons obtenu une première formulation de ce problème comme optimisation d'une fonction sous contraintes linéaires sur des variables binaires ainsi qu'un certain nombre de variantes de cette formulation pour plusieurs compromis efficacité de la recherche - taille de la formulation. Cependant les premières expériences montrent que les outils génériques d'optimisation que nous avons utilisés (Cplex) ne permettent pas de traiter directement le problème efficacement. La formulation du problème nécessite d'être retravaillée, avec introduction de connaissances supplémentaires propres au domaine. La difficulté majeure semble la difficulté à décomposer le problème initial en composantes faiblement couplées.

6.1.4. Inférence grammaticale : application à la bioinformatique

Participants : François Coste, Jacques Nicolas, Daniel Fredouille, Aurélien Leroux, Ingrid Jacquemin, Yoann Mescam.

Notre apport en découverte de motifs repose sur la considération des ensembles de séquences comme des langages produits par des générateurs (machines) qu'il s'agit d'identifier (apprendre). L'hypothèse de base sur laquelle se fonde l'approche, comme dans de nombreux travaux en apprentissage automatique, est que la machine la plus probable est la plus petite. Plus précisément, dans notre cas, l'hypothèse est que le mécanisme de reconnaissance d'un site ou de construction d'une famille de séquences est codé de manière optimale, c'est à dire que sa complexité est réduite autant que possible, minimisant les moyens dépensés par une cellule où les contraintes d'encombrement et la profusion des mécanismes en compétition sont particulièrement fortes. Ce type d'apprentissage à partir de séquences est l'objet d'étude de l'inférence grammaticale.

Nous nous intéressons plus particulièrement aux travaux tendant à renforcer l'applicabilité pratique des techniques d'inférence. Notre objectif est de démontrer que, sur des corpus réels de données biologiques, et moyennant un certain nombre de recherches, les résultats de l'inférence grammaticale sont transférables et pertinents.

6.1.4.1. Incorporation de connaissances biologiques

Dans le cadre du stage de DEA Génomique et Informatique de H. Thomas, nous avons étudié l'introduction de connaissances biologiques pour obtenir des heuristiques performantes d'apprentissage d'automates classifieurs sur des séquences protéiques. Une adaptation de la meilleure heuristique connue dans le domaine par incorporation d'une matrice de substitution entre acides aminés a été réalisée et a été testée sur une nouvelle base d'exemples d'apprentissage composées de séquences de protéines MIP. Ce travail constitue un premier pas vers l'incorporation de connaissances biologiques au cours de l'apprentissage et est poursuivi dans le cadre de la thèse d'A. Leroux. Un premier axe de travail consiste à étudier les conséquences de l'introduction d'un

ordre sur le vocabulaire sur lequel sont construits les mots du langage. Ainsi, l'ensemble des acides aminés possède des propriétés physico-chimiques variées qui permettent de définir plusieurs hiérarchies sur celui-ci. Un deuxième axe consiste à étudier la fusion de transitions plutôt que d'états lors de l'inférence d'automates, afin de permettre d'inférer des automates plus complexes au niveau des mécanismes de transition (transition sur des mots ou des langages simples). D'autres approches permettant de restreindre l'espace de recherche à partir de connaissances d'un expert ont également été développées dans le cadre de la thèse de D. Fredouille.

6.1.4.2. Classe des langages acceptés

L'état de l'art correspond à l'inférence de grammaires régulières. Une idée d'extension de la portée des méthodes d'inférence régulière consiste à travailler à un niveau plus abstrait, sur les chaînes de production d'une grammaire universelle de plus haut niveau (langages de Szilard). On peut ainsi contrôler un langage algébrique, comme ceux observés dans les phénomènes d'appariement de nucléotides ou d'acides aminés, par un langage régulier. I. Jacquemin travaille ainsi sur la prédiction de ponts disulfure dans le cadre de sa thèse. Le problème a été modélisé par une grammaire algébrique universelle et de premières études expérimentales ont permis de souligner les différentes difficultés de la méthode : génération de contre-exemples, traitement du bruit.

La thèse de Y. Mescam porte également sur l'apprentissage de langages adaptés à la description des motifs généralement rencontrés dans les protéines. Cette représentation s'inspire des grammaires à variable de chaîne de Searls qui permettent de modéliser des langages contextuels. L'inférence de telles expressions ne peut raisonnablement pas être réalisée de manière exacte, nous ferons donc appel à des algorithmes de type évolutionniste, en coopération avec le SIB à Genève. Nous étudions actuellement le problème des motifs doubles (dyades) dont les sous-parties sont reliées par un morphisme.

6.2. Analyse et identification de systèmes dynamiques

Participants : Michel Le Borgne, Israël-César Lerman, Jacques Nicolas, Anne Siegel, Basavanneppa Tallur, Andre Floeter, Yves Bastide, Elodie Retout.

Nous avons pour objectif de démontrer l'identification d'un système complet de gènes impliqués dans une voie métabolique particulière. On cherchera également à développer un environnement pour l'analyse de la dynamique des systèmes dynamiques en biologie moléculaire (interactions protéines/gènes, protéines/protéines ...). Cette approche repose sur la synthèse sous forme de modèle qualitatif des connaissances disponibles dans les articles et sur l'identification des modèles par différentes techniques (apprentissage, analyse de données).

Ce thème correspond à des demandes fortes du point de vue de la biologie et à un champ de recherches encore émergeant auquel nous souhaitons contribuer. Il s'agit du traitement des données d'expression des gènes, dont le but ultime est d'aboutir à des modèles de systèmes dynamiques décrivant des interactions protéines/protéines, protéines/gènes ..., impliquées dans les mécanismes biologiques. La production de données d'expressions est un thème majeur de la Génopole Ouest et permet d'intégrer les connaissances individuelles sur les gènes avec une vue globale sur le comportement d'ensembles de gènes et protéines dans un tissu et des conditions donnés.

Le premier niveau d'analyse de ce type de données est clairement un problème de classification (les gènes corégulés sont susceptibles d'exhiber des comportements corrélés : comment définir cette corrélation au mieux pour regrouper ensemble ces familles de gènes ?) et conduit essentiellement à effectuer un diagnostic des gènes impliqués dans une situation donnée (pathologie en particulier). Nos recherches en classification visent en particulier à adapter la méthode AVL développée par I.-C. Lerman aux problèmes particuliers que posent ce type de données : classification d'objets avec variables structurées et classification de grands ensembles (faibles données). Le second objectif, beaucoup plus ambitieux, est de mettre à disposition des biologistes, un outil de travail leur permettant d'appréhender l'ensemble des interactions moléculaires d'intérêt pour leur sujet d'étude, dans une démarche holistique, afin de pouvoir émettre de nouvelles hypothèses et les plans d'expérience pour les tester. Ceci suppose un travail de modélisation, le développement d'environnements de simulation, ainsi que, et c'est l'objectif sur lequel nous souhaitons approfondir nos recherches, la conception de méthodes et d'outils d'identification des modèles (au sens de l'automatique, i.e. d'apprentissage).

6.2.1. Classification

Participants : Israël-César Lerman, Jacques Nicolas, Basavanneppa Tallur, André Floeter, Yves Bastide.

Le contexte général où se situent nos travaux est celui d'une interaction entre, d'une part, une approche de classification non métrique, combinatoire et statistique et, d'autre part, un ensemble de problèmes algorithmiques fondamentaux qui se présentent dans l'analyse de données complexes. L'aspect classification comprend aussi bien la classification non supervisée par Analyse de la Vraisemblance du Lien (AVL, programme CHAVL) que celle supervisée qui relève de la discrimination par arbres de décision.

6.2.1.1. Analyse de la méthode AVL

La méthode AVL (Analyse de la Vraisemblance des Liens) est peut être davantage connue pour la classification de l'ensemble V des variables descriptives que pour celle d'un ensemble O d'objets ou C de catégories décrits au moyen de V. Cependant cette méthode permet avec la même rigueur conceptuelle d'élaborer une classification ascendante hiérarchique sur O (respectivement sur C) et de fournir des coefficients d'« explication » compte tenu de l'organisation de V.

Nous avons repris avec la collaboration de Ph. Peter (Ecole Polytechnique de Nantes) l'analyse conceptuelle et expérimentale de la construction d'un indice de similarité probabiliste entre objets décrits par des variables de types quelconques. Ceci, afin de la comparer le plus finement possible, avec une approche due à W.D. Goodall qui conduit également à un indice probabiliste.

Notre méthode s'avère essentiellement distincte : plus souple, très générale et surtout, tenant étroitement compte de la sémantique des variables, notamment dans le cas qualitatif[33][17].

6.2.1.2. Critères linéaires de validation d'une classification

Nous avions établi un critère de validation d'une classification, bien fondé sur les plans formel et statistique. Ce critère confronte une partition donnée de l'ensemble E décrit avec une information de nature ordinale ou numérique quant aux ressemblances entre éléments de l'ensemble E. L'expression de ce critère est essentiellement quadratique par rapport à la taille de E. Ce qui peut limiter son usage dans le cas de « très grosses données » tel qu'il s'en présente dans la Fouille de Données (Data Mining).

Suite à l'extension de l'algorithme des K-MEANS dans le cas de larges ensembles de données où les variables sont numériques ou qualitatives nominales [58], nous avons bâti une expression linéarisée de notre critère qui s'applique dans le cas d'un ensemble d'objets décrits par des attributs numériques, booléens ou qualitatifs nominaux. Les résultats expérimentaux sur des données difficiles de quelques dizaines de milliers de points, ont dès lors montré le grand intérêt de ce critère. Cette recherche est menée en collaboration avec J.P. Costa (Université de Porto)[22][26].

6.2.1.3. Classification hiérarchique de gros ensembles libre ou sous contrainte de contiguïté

Il est maintenant bien admis et depuis longtemps que la technique de recherche des plus proches voisins réciproques est cruciale pour la conception d'algorithmes de construction ascendante hiérarchique d'arbres de classification sur de « gros » ensembles. La situation spécifique considérée et qui se retrouve dans nombre d'applications est celle où, pour la formation des classes, une contrainte de contiguïté doit être respectée. On suppose de plus que le nombre d'objets contigus à un objet donné reste limité par une constante fixée à l'avance. C'est typiquement la situation pour la classification des pixels d'une image numérisée. K. Bachar avait, notamment dans le cadre de sa thèse [37], élaboré et analysé sur les plans théorique et expérimental, un algorithme CAHCVR (Classification Ascendante Hiérarchique sous Contraintes utilisant les Voisins Réciproques). On démontre et on vérifie dans la pratique que la complexité moyenne en temps de calcul devient linéaire, au lieu de quadratique dans le cas général, en fonction du nombre d'objets, ce qui est optimal.

Dans ces conditions, nous avons avec la collaboration de K. Bachar complètement repris l'étude algorithmique à la fois sur les plans théorique, logiciel et expérimental. Deux types de critères pour l'émergence des classes sont considérés et comparés. Il s'agit d'une part du critère de l'inertie expliquée et d'autre part, de la famille de critères de dissmilarité « informationnelle » issue de la méthode AVL de la vraisembance des liens. D'autre part et dans chacun des cas, l'algorithmique procède par agrégation multiple des paires de classes

réalisant « en même temps » la plus grande proximité. Par rapport à l'agrégation binaire, notre approche respecte « mieux » les dernières dissimilarités observées entre classes. Elle permet d'autre part, un gain non négligeable de performance face au traitement de données de grande taille. Il s'est avéré expérimentalement, relativement aux critères issus de l'AVL que l'arbre de classification sous contrainte de contiguïté ne présentait pour ainsi dire, pas d'inversions ; alors que des inversions pouvaient « facilement » se produire dans le cas du critère de l'inertie expliquée. À cet égard un résultat théorique vient d'être établi. Cette recherche [25] est actuellement dans une phase très active.

6.2.1.4. Qualité des règles d'association en fouille des données

Un objectif fondamental de la fouille des données pour l'ECD (Extraction de Connaissances à partir de Données) consiste à découvrir et à valuer numériquement des règles non symétriques du type « si A est vrai alors B a une propension à l'être ». Une telle valuation s'obtient au moyen d'un indice d' « implication ». L'indice probabiliste implicatif usuel de vraisemblance du lien évaluant de façon intrinsèque une règle d'association a été introduit dans la thèse de Régis Gras [54] à partir de l'indice de similarité probabiliste symétrique, considéré dans l'approche « vraisemblance du lien ». Cet indice rencontre significativement le concept d'intérêt d'une règle. Cependant, compte tenu de son caractère local, il n'est plus discriminant pour la comparaison mutuelle de plusieurs règles dès lors que le nombre d'observations augmente suffisamment. Nous avons dans ces conditions montré et validé l'extension discriminante de ce type d'indice. Cette dernière s'effectue dans le contexte d'un ensemble de règles d'association d'une façon qui s'inspire de la méthode de construction d'un indice de similarité probabiliste discriminant dans la classification hiérarchique AVL. Nous avons, conformément à un protocole expérimental adéquat où le nombre d'objets augmente, étudié le comportement du nouvel indice probabiliste préalablement normalisé. Ce dernier s'avère toujours discriminant. D'autre part, il tend vers une valeur intuitivement raisonable compte tenu du contexte des autres règles d'association. Cette valeur limite nous permettra de situer l'intérêt relatif d'une règle. Ce travail a été conduit en collaboration avec J. Azé de l'Équipe Inférence et Apprentissage du LRI (Univ. Paris Sud). Il s'inscrit dans le groupe de travail GafoQualité de l'action spécifique GafoDonnées du département Stic du CNRS [32].

6.2.1.5. Classification prédictive de protéines MIP

Nous avons une bonne expérience en classification hiérarchique des données par la méthodologie AVL que nous avons adaptée pour la classification des séquences, alignées et non alignées [88], des protéines. L'indice de similarité entre les séquences protéiques qui est construit conformément aux principe AVL, c'est-à-dire sous la forme d'une vraisemblance, utilise une matrice de similarité entre acide aminés. Le critère basé sur la préordonnance introduit dans le cadre de la méthode AVL, permet d'évaluer la cohérence d'une partition obtenue à un niveau donné de la classification hiérarchique par rapport à la mesure de similarité. Nous avons eu l'idée d'utiliser ce critère pour « régler » les paramètres, y compris la matrice de similarité entre acides aminés, pour la prédiction de la fonction des protéines. Les expériences menées sur la famille des protéines MIP ont été très significatives[52] [11]. D'autres voies sont actuellement explorées pour la prédiction fonctionnelle en utilisant les méthodes de classification supervisée conduisant à un arbre de décision telle que la méthode CART.

6.2.1.6. Recherche de variants génétiques discriminants dans l'homéostasie du fer

Cette recherche s'inscrit dans le cadre d'une collaboration entre Jean Mosser et I.-C. Lerman au sein du projet Fer de l'UMR-CNRS 6061 (Génétique et développement). Le problème général consiste à déterminer des profils génétiques responsables ou accompagnant l'hémochromatose (surcharge en fer). À cet égard, un premier ensemble d'apprentissage ou échantillon (au sens statistique du terme) est en cours de constitution. Il sera formé de 1000 sujets bretons normaux. À cet échantillon sera confronté le moment venu, un ensemble de même taille formé de personnes carencées. Sur chaque individu de l'ensemble d'apprentissage des paramètres quantitatifs tels que le fer sérique, le coefficient de saturation de la transferrine et la ferritine, sont mesurés. D'autre part, il y a lieu d'inférer les haplotypes des différents individus relativement à 9 gènes et à 6 ou 7 sites polymorphiques par gène, chacun des sites pouvant être occupé par l'un d'entre deux nucléotides. L'étape actuelle est une étape de constitution des données. À ce jour, 7 SNPs de la transferrine ont été génotypés chez

plus de 400 individus bretons non apparentés pour lesquels on dispose à la fois du bilan martial complet et des données génétiques concernant la mutation C282Y du gène HFE1 fortement impliqué dans la pathologie de l'hémochromatose. Le travail a été mené par Valérie Dehais (Assistante hospitalo-universitaire attachée au serrvice de Génétique et Endocrinologie de Génétique Moléculaire du CHU (Professeur Véronique David)) sous la direction de Anne Marie Jouanolle (Praticienne hospitalière) et de Jean Mosser. L'étape suivante consistera à déterminer une stratégie optimale de classification et d'analyse combinatoire des données pour découvrir des régréssions intéressantes.

6.2.1.7. Aide à l'identification de réseaux métaboliques et de régulation

L'étude de données transcriptomiques fait l'objet d'une collaboration avec l'université de Potsdam et d'une thèse en cotutelle, codirigée par J. Nicolas et T. Schaub (A. Floeter). Il s'agit d'identifier cette fois des chemins métaboliques. Les données de spectrométrie de masse (GC -Gaz Chromatography- et HPLC -High Performance Liquid Chromatography-) sont fournies par le Max Planck Institute de Berlin. Les techniques étudiées sont les arbres de décision et la programmation logique inductive.

6.2.2. Analyse de textes en langage naturel

Participants : Michel Le Borgne, Israël-César Lerman, Jacques Nicolas.

Une autre source de données est fournie par les publications scientifiques associées au dépôt de séquences. Nous participons au projet Caderige (appel d'offre inter-EPST Bio-Informatique), qui vise à filtrer les descriptions potentielles d'interaction géniques dans la masse des résumés Medline associés aux banques. La batterie d'outils d'analyse du langage naturel utilisée fait essentiellement appel en ce qui nous concerne d'une part à la classification et d'autre part à la programmation logique inductive, en collaboration avec P. Sébillot, M. Rossignol et V. Claveau (projet TexMex). Dans les expériences que nous menons, les mots du corpus sont lemmatisés et étiquetés catégoriellement (à l'aide de l'étiqueteur de Brill). On cherche à détecter automatiquement les groupes de mots marqueurs d'un thème (par exemple description des conditions d'expérimentation, interaction, etc). Les groupes de mots, liés au corpus d'étude, ne sont pas connus a priori mais sont directement appris sur les données textuelles disponibles. Une zone de texte est ensuite reconnue comme abordant tel ou tel thème si elle contient au moins deux des mots appartenant au groupe de mots correspondant à ce ou ces thèmes.

Cette année, nous avons mené un travail de mise au point d'une méthode d'apprentissage pour la sélection de fragments de textes mentionnant une interaction entre gènes. Cette étude, effectuée dans le cadre d'un DEA (P. Tchienhom), a reposé sur l'utilisation de techniques de programmation logique inductive (PLI). À partir d'un corpus de 2209 résumés extraits de Medline, nettoyé, lemmatisé, étiqueté, nous avons préparé un échantillon d'exemples positifs et négatifs. Le programme de PLI Aleph a été utilisé (A. Srinivasan), ce qui supposait la mise au point d'une théorie du domaine et de prédicats de description pertinents. En particulier, nous avons intégré dans la théorie des regroupements en classe de mots en utilisant des radicaux communs et grâce à une classification AVL utilisant la distance de Levesthein comme indice de dissimilarité des chaînes. Le taux de précision obtenu est de 66,7 % et le taux de rappel varie de 74 à 80,5% si on utilise les classes de mots.

6.2.3. Maîtriser la complexité des modèles

Participants : Michel Le Borgne, Anne Siegel, Elodie Retout.

La complexité des modèles d'interaction figure parmi les principaux obstacles à surmonter, qu'il s'agisse de simulation, d'analyse ou d'identification. Notre but à terme n'est pas de nous limiter à un seul type de modèles, mais au contraire une panoplie de modèles plus ou moins abstraits et donc plus ou moins explicites. Comme nous l'avons détaillé dans la partie fondements scientifiques, un certain nombre de publications ont déjà montré l'intérêt de modèles qualitatifs ou semi-qualitatifs, bien adaptés par rapport à la faible précision des technologies actuelles ainsi que celle des données textuelles. On ne cherche pas à modéliser les interactions moléculaires fines (modèles de connaissance) mais bien les comportements du système. Ceci nous semble d'autant plus raisonnable que les ensembles de gènes forment en général des systèmes dynamiques

extrêmement robustes, capable de fonctionner correctement malgré des variations importantes de certains constituants.

Nous avons donc repris un modèle d'interactions géniques déjà utilisé par plusieurs équipes : le modèle à base d'équations différentielles constantes par morceaux. Ce modèle peut-être approximé par un modèle qualitatif qui n'est autre qu'un automate non déterministe à espace d'états fini. Les équipes travaillant sur ce modèle ont porté leurs efforts sur les rapports entre propriétés du modèle qualitatif et celles du modèle différentiel (école « belge ») ou la simulation de l'automate non déterministe (projet Helix). Nous avons choisi d'utiliser notre savoir faire en représentation des systèmes à évènements discrets pour obtenir une représentation par fonctions logiques de l'automate. Cette représentation devant permettre l'étude de propriétés et compléter ainsi les simulations. L'efficacité de telles représentations, utilisées depuis longtemps en vérification de circuits, devrait permettre de modéliser de plus grands réseaux de gènes que ceux qu'il est possible d'étudier en simulation. Il reste à démontrer l'utilisation concrète de ces outils dans le cadre d'une collaboration avec des biologistes.

Un autre élément de maîtrise de la complexité est de cibler la modélisation sur les ensembles de gènes impliqués dans une voie métabolique particulière, ce qui nous semble d'ailleurs la seule voie d'étude raisonnable, contrairement à une tendance encore active qui a l'ambition de monitorer l'ensemble des gènes d'un organisme (cas de la levure) ! Il y a en effet un énorme pas entre la capacité d'observation (on arrive à faire tenir l'ensemble des gènes de la levure sur une même puce à ADN) et la capacité de modélisation (actuellement une dizaine de gènes) ou d'identification de ces modèles. Nous avons ainsi initié une collaboration avec le laboratoire de génétique animale de l'Inra Rennes dans le cadre d'une étude systémique sur la lipogénèse chez le poulet. Comme dans de nombreux problèmes bioinformatiques, la phase d'expression des besoins est longue. Des difficultés sont apparues dans l'interprétation de certains résultats expérimentaux ainsi que d'une facon plus générale dans la compréhension du réseau des interactions qui sont ici à la fois de nature métabolique et génique.

Notre activité est donc maintenant orientée vers l'étude de tels réseaux mixtes, orientation qui avait été préparée par des études préliminaires lors de deux stages de maîtrise. Ceci comporte la définition d'un langage de description de ces réseaux et l'élaboration d'un outil convivial permettant l'exploration de certaines propriétés du réseau par un biologiste. Nous étudions en parallèle la modélisation qualitative de ces réseaux et les techniques permettant de prédire leur comportement ou d'expliquer certaines observations. Le but final étant d'intégrer ces techniques dans l'outil en cours de développement.

Nous espérons également collaborer avec le laboratoire Inra Scribe de Rennes sur la gestion du stress chez la truite, avec une approche similaire.

6.2.4. Identification des modèles

Participants : Michel Le Borgne, Anne Siegel, Aurelien Leroux.

Vouloir apprendre un modèle complexe ab initio est une démarche vouée à l'échec. Les connaissances sur les interactions sont déjà nombreuses et reflétées aussi bien dans les manuels classiques que dispersées dans des publications et dans les bases de données publiques. Ces connaissances sont toujours parcellaires, non formalisées et isolées du contexte des autres interactions. On ne peut espérer complètement automatiser ce recueil de connaissances, mais il reste important d'assister ce processus. Nous y participons de deux manières : tout d'abord par la mise au point de modèles de bases de données adaptés au problème, facilitant la mise à jour à partir des banques publiques et le suivi de la qualité des données récoltées. Ceci fait l'objet d'une collaboration avec L. Berti, du projet TexMex. Ensuite par le relevé de zones potentiellement intéressantes dans les textes des publications (voir le paragraphe précédent sur la classification).

Il s'agit ensuite de construire les modèles à partir des bases de données et des données d'expérimentation. Nous travaillons à la formalisation de problèmes d'apprentissage bien délimités sur ces données, tout particulièrement en inférence grammaticale. Une possibilité est de s'intéresser à une propriété particulière sur le réseau. Ainsi, nous avons prévu d'étudier l'accessibilité des états à partir des suites observées par inférence d'automates (Thèse d'A. Leroux).

6.3. Parallélisme

Participants : Rumen Andonov, Dominique Lavenier, Frédéric Raimbault, Stéphane Guyetant, Mathieu Giraud, Hugues Leroy, Michel Mac Wing, Nicolas Yanev.

Concrètement, l'axe parallélisme se focalise sur trois types d'actions : (1) la conception de machines parallèles spécialisées pour l'exploration des banques de données génomiques ; (2) la parallélisation d'algorithmes complexes issus de la bioinformatique sur des calculateurs parallèles ; (3) le portage d'applications génomiques très coûteuses en calcul sur une grille de calculateurs répartis géographiquement (projet GénoGRID).

6.3.1. Filtrage des données génomiques sur architectures spécialisées

Participants : Dominique Lavenier, Stéphane Guyetant, Mathieu Giraud, Frédéric Raimbault.

L'ensemble des informations génomiques disponibles est stocké dans de gigantesques banques ou des bases de données. Se pose alors le problème de la consultation de ces informations.

Entre la fin des années 1970 et le milieu des années 1980, une recherche très active en bases de données visait à rapprocher le plus possible les traitements de l'endroit où étaient stockées les données. Les diverses approches étudiées affichaient des performances exceptionnellement bonnes car les algorithmes câblés (tris, jointures, filtrages) sur des processus spécialisés étaient fortement optimisés et rendaient très efficaces le traitement des données au vol. À l'époque, la fabrication de ces processeurs au bout du compte peu extensibles était très onéreuse (coûts et surtout délais pour y câbler des algorithmes extrêmement complexes), ce qui, ajouté à l'obligation d'utiliser du matériel non standard (disques, contrôleurs) a entraîné un arrêt total de toute recherche sur ce domaine. Depuis deux ans, l'idée de rapprocher traitements et données refait surface, notamment au travers des projets IDISKs (Berkeley, USA) et Active Disks (CMU, USA). Ces deux projets utilisent les capacités des processeurs intégrés aux contrôleurs de disques pour y télécharger des algorithmes traitant les données en amont, au fur et à mesure qu'elles sont lues, et peuvent les filtrer ou les manipuler avant de les laisser passer vers le bus de la machine. Les gains qui se dégagent de ces approches sont importants et les coûts de mise en oeuvre faibles.

L'idée que nous poursuivons est d'intercaler entre les disques et la mémoire un dispositif matériel de nature très différente des solutions précédentes, puisque nous envisageons l'utilisation de filtres reconfigurables à base de circuits FPGA. À l'instar des autres approches, on peut alors traiter les données à la volée dès la sortie des disques pour ne laisser passer que les informations pertinentes. Cette approche présente deux avantages notables : (i) la reconfigurabilité permet de changer instantanément la nature des filtres, en adaptant la configuration à la nature de l'algorithme et aux caractéristiques (distributions, formats, ...) des données à traiter ; et (ii) on peut également configurer la façon dont les différents étages mis en parallèle sont reliés, ce qui permet cette fois d'optimiser le flux des données offerts aux algorithmes.

Plus précisément, l'architecture spécialisée en cours de définition à l'Irisa repose sur le concept de « disques intelligents » dont le rôle est de filtrer à la volée les informations des banques [23]. Comme la nature des données ou des traitements peut être très variable, les filtres ne peuvent être figés définitivement. C'est pourquoi l'unité de base de l'architecture est composée d'une structure reconfigurable (à base de circuits FPGA) couplée directement à un ou plusieurs disques durs. Plusieurs disques oeuvrent en parallèle pour augmenter le débit global de l'accès aux données.

Au cours de l'année 2002 nous avons concu et réalisé une carte prototype composée d'un disque dur, d'un circuit FPGA (Spartan II, Xilinx), d'un micro-controleur, d'une mémoire et d'une connexion Ethernet. Après validation de ce prototype, il est prévu de fabriquer une petite série de 32 ou 64 cartes qui seront connectées localement par un réseau Ethernet, et sur laquelle les premières expérimentations grandeur nature pourront commencer.

En parallèle nous travaillons sur la mise en oeuvre de deux applications pilotes, la comparaison de séquences d'ADN dans des génomes complets, des banques d'EST, etc., et la recherche de motifs complexes dans des bases de données protéiques et/ou d'ADN. Nous imaginons également un environnement de programmation pour permettre une conception efficace des filtres adaptatifs sur le circuit FPGA [28].

Cet axe de recherche est mené en coopération avec d'autres équipes de recherche de l'Irisa, notamment l'équipe R2D2 pour les aspects matériel (P. Quinton, S. Derrien) et l'équipe TexMex (L. Amsaleg), pour la problématique qui consiste à extraire rapidement des candidats d'une base de données d'images volumineuse à partir de requêtes sur leur contenu.

6.3.2. Alignement de séquences protéiques sur des structures tridimensionnelles

Participants : Rumen Andonov, Dominique Lavenier, Nicolas Yanev.

Les techniques d'analyse fonctionnelle in silico ont pour but d'assigner une fonction aux protéines produits de gènes nouvellement identifiés. Les méthodes les plus employées, comme BLAST ou FASTA, se basent sur des méthodes de comparaison de séquences pour mettre en évidence les relations d'homologie avec des protéines de fonction connue stockées dans des bases de données. Ces méthodes sont rapides, efficaces, et disposent d'un critère objectif pour juger si deux protéines sont homologues. Cependant, elles ne permettent pas de repérer des relations d'homologie entre protéines issues d'organismes éloignés et dont les séquences peuvent avoir considérablement divergé.

La direction de recherche a été inspirée par les résultats récents de l'équipe MIG de l'Inra à Versailles. Le principe de cette méthode repose sur l'alignement d'une séquence sur une structure tridimensionnelle de protéine. Elle fait appel, entre autres, à un algorithme d'optimisation combinatoire permettant d'aligner les séquences sur des modèles qui caractérisent les repliements spaciaux observés. Contrairement aux méthodes de comparaison de séquences, pour lesquelles on dispose d'un algorithme performant pour trouver l'alignement de score optimal (l'algorithme de programmation dynamique), la reconnaissance de repliements est un problème NP-complet pour lequel aucune résolution satisfaisante n'est encore proposée. Deux pistes sont à poursuivre simultanément pour obtenir de meilleures performances de résolution : (1) améliorations de l'algorithme séquentiel existant en utilisant de nouvelles formulations et (2) parallélisation de l'approche séquentielle. Les difficultés de la parallélisation d'un algorithme de type « branch and bound » et « divide and conquer » sont liées surtout à l'équilibrage de charge entre les processeurs. Nous préférons étudier une parallélisation dédiée qui permet seule actuellement d'obtenir une bonne efficacité [27].

Notre dernier résultat dans ce domaine [35] montre que le modèle linéaire en variables binaires pour ce problème s'avère très efficace et donne des résultats fortement encourageants. Nous avons résolu toutes les instances qui nous ont été fournies par les biologistes (i.e. basées sur des données *réelles*) par le logiciel de LP (*Linear Programming*) CPLEX au lieu d'un algorithme *branch&bound* dédié. L'optimum de tous les cas résolus, est atteint dans un sommet (0,1) du polytope sous-jacent, ce qui suggère que le problème (théoriquement NP-complet) pourrait être facilement abordable dans des applications biologiques réelles.

Cet axe de recherche est mené en partenariat avec l'université de Valenciennes.

6.3.3. La programmation dynamique appliquée à la génomique

Participants : Rumen Andonov, Dominique Lavenier.

La programmation dynamique est une des techniques algorithmiques la plus largement utilisée dans le domaine de la bioinformatique. Les difficultés rencontrées lors de la parallélisation de cette approche sur des machines à mémoire distribuée sont typiques des algorithmes réguliers et massivement parallèles : elles sont intrinsèquement liées à la réduction du coût des communications. Une des techniques pour réduire ce coût et contrôler la granularité du calcul en améliorant le rapport calcul/communications est le pavage. L'espace d'itérations, défini par un nid de boucles, est partitionné en unités appelées tuiles. Le pavage optimal consiste à déterminer la taille des tuiles pour minimiser le temps total d'exécution.

Dans [14] nous avons appliqué le pavage semi-oblique à la parallélisation de l'algorithme de Fickett, dit algorithme par bande, rencontré dans le domaine de la bioinformatique. C'est un algorithme de type programmation dynamique utilisé pour l'alignement global de deux séquences similaires de grande taille. Il possède une meilleure complexité que l'algorithme classique de Needlemen et Wunsh et n'avait, à notre connaissance, jamais été parallélisé. Pour certaines instances le pavage semi-oblique se montre 2.5 fois plus rapide que le pavage orthogonal. Notre solution optimale requiert une distribution cyclique par blocs qui s'avère trois fois plus rapide que la distribution non-cyclique utilisée par d'autres auteurs.

Une deuxième application de la technique du pavage concerne une classe où les dépendances apparaissent aussi entre des étapes non-consécutives de la programmation dynamique. Cette classe a été très peu étudiée du point de vue du parallélisme. Nous nous sommes intéressés plus particulièrement au problème de la prédiction de la structure secondaire d'ARN [21]. Nous proposons un nouvel algorithme parallèle pour la classe sous-jacente qui se caractérise par un domaine d'itérations triangulaire et des récurrences non-uniformes. L'approche proposée présente un développement et une extension de la technique du pavage. Nous formulons et résolvons d'une manière analytique le problème pour déterminer la taille optimale des tuiles afin de minimiser le temps total d'exécution sur des machines à mémoire distribuée. Les résultats numériques sur la CRAY T3E confirment la validité du modèle théorique proposé. Ce résultat a été obtenu en collaboration étroite avec l'équipe de la Laguna, Espagne.

6.4. Autres contributions

6.4.1. Modélisation de la fragmentation d'un génome bactérien

Participants : Rumen Andonov, Dominique Lavenier, Nicolas Yanev.

Ces travaux sont menés en coopération avec le laboratoire de microbiologie alimentaire Inra-ENSAR UMR 1055. Les chercheurs de ce laboratoire, en particulier M. Gautier et Y. Leloir, étudient la plasticité du génome de la bactérie pathogène *Staphylococcus Aureus* par génomique comparative. Les génomes de diverses souches de cette bactérie, dont les tailles se situent autour de 2,8 Mb, sont découpés en fragments puis amplifiés par PCR (Polymerase Chain Reaction). À chaque souche correspond un profil (les fragments amplifiés ou non). Il s'agit ensuite de comparer ces profils pour étudier leur disparité - ou leur ressemblance - et étudier l'évolution dans le temps (souches prélevées à différentes périodes) ou dans l'espace (souches prélevées à différents endroits).

Les fragments à amplifier sont identifiés sur la base de deux amorces, une au début et l'autre à la fin. Le problème est de découper le génome de manière optimale par rapport à un jeu de contraintes liées à sa structure et aux techniques d'amplification : les sites favorables aux amorces ne sont pas répartis équitablement le long du génome, la taille des fragments doit être sensiblement identique pour que l'amplification puisse se dérouler dans de bonnes conditions, etc. Nous sommes donc face à un problème typique d'optimisation combinatoire pour lequel l'espace de solutions est gigantesque et où on doit en extraire - si possible - la meilleure.

Trois modélisations ont été étudiées. La première se base sur la programmation dynamique. Le génome est découpé arbitrairement en N fragments avec une zone variable où on autorise les débuts et fins de fragments. Dans cette zone existent - ou non - des amorces sur lesquelles on s'appuie pour construire des solutions qui sont évaluées par une fonction de score prenant en compte la taille des fragments, leurs chevauchements, la compatibilité des amorces les unes avec les autres, etc. La solution est construite étape par étape. Différentes solutions finales peuvent être générées en faisant varier le nombre de fragments N dans une plage raisonnable, par exemple entre 260 et 300 si on vise des tailles de 10 Kb +/- 1Kb.

La seconde modélisation pose le problème sous forme d'un graphe, les sommets représentent les fragments admissibles et les arcs le coût *d'interconnection* de ces fragments. Ce coût est déterminé suivant leur longueur et leur chevauchement. La taille du graphe ainsi obtenu est importante ; par conséquent, trouver une solution par un calcul du plus court chemin est onéreux. Aussi, pour minimiser le temps de calcul, la stratégie appliquée consiste à réduire le graphe par élagages successifs des arcs coûteux tout en vérifiant qu'il existe toujours une solution. Le graphe final contient toutes les solutions optimales par rapport à la fonction de score définie au paragraphe précédent.

Enfin, la dernière modélisation traduit la résolution du problème en terme de programmation linéaire en nombres entiers (*Mixed Integer Programming*). Cette fois le graphe est vu comme un graphe de flot à partir duquel un ensemble de contraintes linéaires est exprimé. La résolution du sytème s'effectue par le *solveur* CPLEX de ILOG déjà utilisé dans l'application d'alignement. La tâche se résume alors à générer cet ensemble de contraintes que l'on obtient aisément à partir de la structure de données de la méthode précédente.

À notre connaissance, il n'existe pas de travaux similaires, pour l'instant, sur ce problème de fragmentation de génomes entiers en vue d'une PCR à grande échelle. Du point de vue précision des solutions, les trois

méthodes sont équivalentes. Les solutions trouvées doivent maintenant être validées par des expérimentations biologiques. Du point de vue performance, la seconde méthode est la plus rapide, suivie par la programmation dynamique, et enfin par l'usage de CPLEX. Cette notion de performance est cependant relative par rapport au temps d'expérimentation biologique qui constitue un travail à la paillasse de plusieurs semaines. À cet égard, l'usage d'un solveur est très souple : la modélisation est rapide et la solution produite automatiquement.

6.4.2. Morphismes itérés, pavages et numération

Participante : Anne Siegel.

Les travaux présentés ici se situent dans la continuité des recherches effectuées par A. Siegel avant son intégration dans l'équipe et ne concernent donc pas la bioinformatique.

Un morphisme itéré [15] constitue une dynamique sur les mots infinis d'un alphabet fini consistant à remplacer chaque lettre par un mot[77]. Il s'agit de modèles naturels, du point de vue de la théorie des langages, des systèmes présentant des propriétés d'autosimilarité. À ce titre ils apparaissent dans diverses situations (pavages quasi-périodiques, codage, théorie des nombres, etc. Voir [13]). Le morphisme est de type Pisot si sa matrice d'incidence a pour valeur propre un nombre de Pisot. On lui associe alors un ensemble compact appelé *fractal de Rauzy*[78][36] par projection sur un plan bien choisi[43] d'une ligne brisée de l'espace décrivant le point fixe du morphisme [29], ou alternativement, par renormalisation dans des plans discrets [20]. Chaque fractal de Rauzy contient les prémices d'une structure autosimilaire : il est recouvert par des morceaux représentant chaque lettre de l'alphabet, chacun se déduisant du fractal de Rauzy global par des transformations géométriques simples. Seule manque à l'autosimilarité la certitude que les morceaux se rencontrent sur un ensemble de mesure nulle. Ceci est vrai dans le cas où le morphisme est unimodulaire et vérifie une condition combinatoire dite de coïncidences[78][36]. On ne connaît aucun morphisme de type Pisot qui ne vérifie pas cette condition. On conjecture qu'il n'en existe pas (voir les survols [18][29]). D'un point de vue dynamique, les fractals de Rauzy sont stables sous l'action d'un déplacement des morceaux par des translations. Dans le cas unimodulaire avec coïncidences, en codant les trajectoires des points du fractal sous l'action de l'échange de morceaux, on retrouve toutes les suites infinies dont le language est celui du point fixe du morphisme [36]. Inversement, à quelques exceptions, toutes les suites ci-dessous correspondent à un unique point du fractal de Rauzy.

6.4.2.1. Pavage périodique

Pour les morphismes sur deux lettres avec coïncidences et dans quelques cas bien déterminés, on sait que le fractal de Rauzy engendre un pavage périodique du plan [78]. De manière générale, on sait seulement que les décalés d'un fractal de Rauzy par un réseau (déduit des propriétés algébriques du morphisme) recouvrent l'espace. Il manque pour avoir un pavage le fait que les décalés s'intersectent sur un ensemble de mesure nulle.

Nous nous sommes concentrés sur la mise au point d'un algorithme effectif permettant de répondre à la question des pavages au cas par cas. On utilise pour celà une méthode de codage du système engendré par un morphisme sous la forme d'un langage rationnel, en utilisant un automate naturellement défini à partir du morphisme (automate des préfixes-suffixes). On peut alors associer au morphisme un système de numération [43] et caractériser, au moyen d'un automate fini, les ambiguités de ce système de numération. En croisant cet automate avec le précédent, on décrit les points qui se retrouvent à l'intersection de deux décalés du fractal de Rauzy. Une étude de l'automate obtenu permet de déterminer si ces points forment un ensemble de mesure nulle. On obtient ainsi un algorithme (implémenté en MuPad, du fait du travail dans des corps de nombres effectué par l'algorithme) qui répond, au cas par cas, à la question des pavages [34].

6.4.2.2. Morphismes non unimodulaires

Dans le cas non unimodulaire, des considérations mathématiques montrent qu'il sera impossible de représenter correctement un morphisme itéré dans un espace euclidien. Nous avons défini un formalisme qui étend la définition des fractals de Rauzy au cas non unimodulaire [19]. On construit ainsi des fractals de Rauzy généralisés, sous-ensembles d'un produit cartésien d'espaces euclidiens (composante réelle semblable aux fractals de Rauzy unimodulaires) et d'extensions finies de corps p-adiques (composante arithmétique qui prend en compte le fait que la matrice d'incidence n'est pas inversible). Les résultats dynamiques relatifs

aux morphismes unimodulaires sont vérifiés par les fractals de Rauzy généralisés ainsi définis. Les méthodes utilisées font appel à de la théorie ergodique combinée avec de la théorie des nombres [19]. La condition de pavage de [34] est encore valable dans le cas non unimodulaire. D'un point de vue mathématique, ceci permet par exemple de prouver que le système substitutif engendré par $1 \longmapsto 2, 2 \longmapsto 3, 3 \longmapsto 11233$ a un spectre purement discret. Aucun exemple de ce type n'avait pu être prouvé jusqu'à présent.

7. Contrats industriels

Nous n'avons pas pour l'instant de contrat en relation avec l'industrie. On sait que la plupart des laboratoires pharmaceutiques sous-traitent l'activité de recherche et le développement de la Génopole Ouest devrait fournir pour cela un cadre propice à la valorisation de nos recherches. Au niveau national, nous avons initié des contacts avec Aureus Pharma, et répondu à un appel d'offre qui n'a pas abouti (I.-C. Lerman).

8. Actions régionales, nationales et internationales

8.1. Projets régionaux

8.1.1. La Génopole Ouest

La Génopole Ouest, huitième génopole sur le territoire, a été créée en janvier 2002 pour une durée probatoire de deux ans. Il s'agit d'un projet stratégique pour la formation supérieure et la recherche dans le domaine des sciences du vivant, de la bioinformatique et pour le développement économique dans les domaines de *la mer*, de *l'agro-alimentaire* et de *la santé*. Une structure de GIS est en création, où les différents EPSTs et universités participant à la Génopole sont représentées (Inra, Inserm, Ifremer, Inria, CNRS, Universités de Rennes, Nantes, Brest et Angers). Le représentant pour l'Inria est C. Labit. J. Nicolas est responsable du domaine Bioinformatique et participe à ce titre aux réunions mensuelles du comité directeur. Il participe également au comité national bioinformatique des génopoles, dirigé par F. Rechenmann.

La création a suscité un certain nombre de visites, et en particulier, trois communications ont été présentées par J. Nicolas à l'Irisa à Rennes aux occasions suivantes :

- Visite de R.G. Schwartzenberg (février) ; *Génopole Ouest et Bioinformatique*.
- Visite de P. Tambourin et J. Haiech (mars) ; *Présentation du domaine Bioinformatique de la Génopole Ouest*.
- Visite de C. Roucairol (mai). *Présentation du pôle Bioinformatique*.

8.1.2. La plateforme Bioinformatique

Participants : Cynthia Alland, Esther Kaboré, Hugues Leroy, Michel Mac Wing, Emmanuelle Morin, Jacques Nicolas.

En s'appuyant sur le contrat de plan État-Région et en regroupant les compétences sur certains sites, 5 plateaux techniques ont été définis dans le cadre de la Génopole Ouest. Il s'agit essentiellement :

1. d'une plate-forme Puces à ADN, à Nantes (IFR26, Unité Inserm 533) ;
2. d'une plate-forme Protéome, à Rennes (Inserm U435) ;
3. d'une plate-forme Génotypage-Séquençage à Roscoff (CNRS, Station de Biologie Marine) et au Rheu (Inra UMR 118) ;
4. d'une plate-forme Exploration fonctionnelle à Rennes (IFR 91,97,98) et à Nantes en ce qui concerne le transfert de gènes avec vecteurs viraux (IFR 26) ;
5. d'une plate-forme Bioinformatique, plate-forme « virtuelle » organisée autour d'un calculateur parallèle installé à l'Irisa (SunFire 40 processeurs), avec des centres secondaires sur Brest, Nantes, Roscoff et Angers.

Avec O. Collin de Roscoff, H. Leroy est responsable, en attendant un ingénieur spécialement affecté à cette tâche, du comité de pilotage de la plate-forme. Des actions de formation sont menées (formation au package GCG, etc) et deux ingénieurs en CDD sont récemment venus renforcer les effectifs pour assurer le bon fonctionnement de la plate-forme sur le plan de la gestion des données et des logiciels, favorisant ainsi les possibilités de coopération inter-disciplinaires.

8.1.3. Agenae

Participants : Elodie Retout, Jacques Nicolas.

Le programme Agenae (Analyse du GENome des Animaux d'Elevage) est un programme national Inra qui a pour ambition de développer des démarches génériques et des actions de recherche finalisées dans le domaine de la génomique animale. Il vise au sein de plusieurs espèces d'animaux d'élevage (porc, poule, truite, vache) à identifier la partie exprimée du génome, à développer la cartographie des génomes entiers et à étudier la diversité génétique dans les populations animales. Le programme de recherche va être piloté dans le cadre d'un Groupement d'Intérêt Scientifique (GIS) constitué pour 5 ans. Il associe des organismes publics de recherche (Inra et Cirad) et des structures professionnelles (Apis-Gene, Cipa). Au niveau international, un partenaire privilégié est l'ARS (Agricultural Research Service) américain qui développe un projet comparable.

Les transcriptomes de deux espèces, la truite et la poule, sont étudiés à Rennes. Au sein de ce projet, E. Retout, a pour rôle, en travaillant dans le projet Symbiose, de participer à la construction du système d'information SIGENA, base de données de séquences et d'expression, de développer des outils de gestion des clones et d'analyse du transcriptome, et d'effectuer de la cartographie comparative entre espèces apparentées.

Cette année, en travaillant au sein du projet Symbiose, E. Retout s'est particulièrement attachée à réaliser les missions de développement suivantes :

- soumission des séquences privées aux bases de données publiques ;
- gestion de la redondance des banques ;
- analyse statistique de la qualité des séquences.

E. Retout assure aussi la cohérence des développements du projet Stressgenes par rapport à ceux d'Agenae. Les perspectives concernent la mise en place d'une base de données spécialisée poulet, qui pourra servir de support à des recherches communes sur l'analyse et l'intégration des données d'expression au sein de réseaux d'interaction géniques et la caractérisation des zones régulatrices de familles de gènes impliquées dans des (sous-)métabolismes particuliers.

8.2. Projets nationaux

8.2.1. Projet GénoGRID

Participants : Dominique Lavenier, Hugues Leroy, Rumen Andonov, Frédéric Raimbault, Michel Mac Wing.

Il s'agit d'une action concertée incitative nationale dans le cadre de l'appel d'offre Globalisation des Ressouces Informatiques et des Données (ACI GRID) lancé par le ministère, dont D. Lavenier est le coordinateur. Cette ACI a pour objectif de mettre en place un portail par lequel des chercheurs en biologie peuvent accéder à des ressources en calcul réparties géographiquement [24]. Les partenaires incluent divers laboratoires du Grand Ouest (Inra, Ifremer, Station Biologique de Roscoff), l'ABISS à Rouen, le LAMIH à Valenciennes, le LIH au Havre et le LIFL à Lille.

Ainsi, ce projet correspond à une action directement inspirée par la structure de la Génopole Ouest, qui lui fournit un terrain d'expérimentation particulièrement propice. Il est suivi par Hugues Leroy, à 50% de son temps et un ingénieur expert, embauché à cet effet (Michel Mac Wing). Il s'effectue en étroite collaboration avec d'autres partenaires, en particulier l'équipe Adept de l'Irisa dont le logiciel PARadis, développé au sein de cette équipe, servira d'ossature au projet GénoGRID.

Sous l'angle des applications génomiques, ce projet se donne comme objectif de proposer des services non standards en matière d'analyse de génomes, en particulier des services extrêmement gourmands en puissance de calcul. Par non standard, nous entendons des services en ligne (accessibles à partir d'une interface Web) non proposés par les serveurs usuels, tel le serveur Infobiogen, par exemple.

La grille est composée de ressources hétérogènes réparties géographiquement dans divers laboratoires du Grand Ouest (Inra, Ifremer, Station Biologique de Roscoff, PCIO : Pôle de Calcul Intensif de l'Ouest, Irisa, etc). Sur cette grille cohabitent des serveurs classiques, des machines parallèles puissantes, des clusters de PC et des machines spécialisées. Les données (i.e. les banques de séquences, les génomes, etc) sont plus ou moins réparties sur les différents noeuds de la grille, tout en sachant que les principales banques sont en général disponibles localement dans chaque centre de recherche, donc sur la majorité des noeuds de la grille.

L'un des défis est de pouvoir accéder à cet ensemble de ressources de la manière la plus transparente possible. L'outil indispensable de ce projet est un portail Internet qui doit fournir à une communauté homogène d'utilisateurs (biologistes et bioinformaticiens) un moyen d'accès simple aux ressources de calcul.

Notre objectif n'est pas de couvrir tous les domaines de recherche couverts par le *Grid Computing,* mais de proposer des solutions pratiques pour les questions suivantes : comment régler l'accès transparent aux ressources distribuées et gérées par des organisations distinctes n'ayant pas forcément les mêmes procédures d'exploitation, comment assurer la migration des résultats tout en garantissant la confidentialité, comment régler l'allocation des ressources et le partage de charge, etc.

L'année 2002 a vu le démarrage de ce projet. Plusieurs actions ont été menées en parallèle :

- la mise en place d'un portail Internet sécurisé : elle se base sur une procédure d'authentification qui attribue des signatures électroniques (certificats) aux utilisateurs.
- la mise en place d'un répertoire LDAP : il mémorise pour chaque utilisateur ses prérogatives (droits d'accès aux applications, aux ressources, etc), pour chaque application la liste des sites où elle peut s'exécuter, ses paramètres, etc.
- la parallélisation d'une première application : la comparaison de banques de séquences conséquentes, suivi d'un filtrage pour ne récupérer que l'information pertinente.
- la mise au point du logiciel Eden, logiciel qui assure la répartition et la synchronisation des tâches déployées sur la grille (en coopération avec l'équipe ADEPT).

La deuxième année du projet sera consacrée à la *gridification* d'au moins deux applications sur une version restreinte de la grille. Il est prévu une première démonstration de faisabilité courant 2003.

8.2.2. Architectures reconfigurables

Participants : Dominique Lavenier, Mathieu Giraud, Stéphane Guyetant, Frédéric Raimbault.

Ce contrat s'effectue dans le cadre de l'action inter-EPST Bioinformatique. Il concerne la mise au point d'architectures parallèles reconfigurables pour l'extraction des données génomiques. Le projet vise l'extraction rapide, et par le contenu, des données génomiques emmagasinées dans les banques et les bases de données. *Par le contenu* indique que la recherche porte sur l'information brute, et non sur les annotations pouvant y faire référence. Par exemple, extraire des séquences sur la base d'un alignement significatif, ou sur la base d'un motif exprimé à l'aide d'une expression régulière, est une tâche qui s'effectue essentiellement sur le texte des séquences, et non sur les annotations.

8.2.3. Caderige

Participants : Michel Le Borgne, Jacques Nicolas.

Le projet Caderige (catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques) est un pré-contrat de un an obtenu en 2000, qui a été renouvelé pour deux ans en octobre 2001. Cette action inter-EPST Bioinformatique CNRS, Inserm, Inra, Inria, Ministère de la Recherche regroupe, outre des membres des projets TexMex et Symbiose, des membres des laboratoires Leibniz de l'Imag, du LIPN, du

LRI, et de deux laboratoires Inra : MIG et Inra-Ensar. Il est dirigé par G. Bisson, de l'IMAG à Grenoble, en coopération avec P. Sébillot.

Il vise la catégorisation automatique de documents pour l'extraction de réseaux d'interactions géniques : son objectif est de filtrer, dans des bases textuelles de bioinformatique telles que MedLine, les textes parlant spécifiquement d'interactions géniques, et d'extraire de ces textes des réseaux de telles interactions.

La participation de notre équipe concerne d'une part la détection des zones de textes susceptibles de contenir une interaction et, d'autre part, la modélisation de l'interaction repérée. Le premier point a fait l'objet d'un stage de DEA (P. Tchienhom), en collaboration avec P. Sébillot du projet TexMex et conduit à une caractérisation par programmation logique inductive des phrases pertinentes vis à vis d'une description d'interaction. Concernant le second point, notre travail consiste à extraire, en relation avec le laboratoire de génétique animale de l'Inra de Rennes, les fragments de résumés d'articles scientifiques consacrés aux 200 gènes actuellement connus du métabolisme des lipides du foie. Ce métabolisme, bien que limité, reste suffisamment complexe pour constituer un modèle d'étude des interactions géniques chez les eucaryotes supérieurs, que nous utiliserons ensuite dans notre axe sur l'analyse des systèmes dynamiques d'interaction.

8.3. Projet européen : StressGenes

Participants : Yves Bastide, Jacques Nicolas.

En novembre 2001 a débuté le contrat européen STRESSGENES (n° Q5RS-2001-02211, *Quality of Life and Management of Living Resources Area 5.1.2*), consacré à l'étude de gènes impliqués dans la résistance au stress chez les poissons. Les partenaires incluent l'Inra (Scribe) à Rennes, porteur du projet, et les universités d'Aberdeen, de Galway, de Liverpool et d'Uppsala. Les études concernent un organisme modèle pour la pisciculture, la truite arc-en-ciel. L'approche poursuivie est celle de la génomique fonctionnelle, via la fabrication de puces à ADN (*macro-* et *micro-arrays*) dédiées.

Notre équipe doit proposer aux autres partenaires du projet les techniques et outils informatiques pour

1. la gestion des données issues des expérimentations,
2. la recherche et l'extraction de données pertinentes depuis les bases de données publiques,
3. l'exploitation conjointe des données issues de ces deux sources.

Après avoir mis en place une liste de discussion et un espace de fichiers partagé, nous avons réalisé un cahier informatique de laboratoire, qui permet aux biologistes de saisir au fur et à mesure leurs données sous forme électronique. Ce cahier comprend les descriptions des différentes expériences réalisées, avec les protocoles suivis, les matériaux biologiques utilisés, et les résultats trouvés. Comme le cahier manuscrit d'usage dans les laboratoires, il impose qu'une expérience finalisée ne peut plus être modifiée. La deuxième étape, en cours, concerne la centralisation des données issues du séquençage d'ADN dans une base de données commune aux différents partenaires. Elle implique aussi leur recoupement entre elles et avec les banques de données publiques de gènes, protéines et références bibliographiques.

D'autres données, issues de l'utilisation des puces à ADN qui vint être fabriquées, devront ensuite être intégrées en suivant la norme MIAME [38] et le modèle MAGE [10].

La phase finale sera alors pour nous un transfert des techniques de classification et de découverte de motifs étudiées dans l'équipe au cas des données d'expression du génome de la truite.

Le rapport d'étape de la première année du contrat est disponible sur le Web [31].

[10] MIAME est la liste minimale d'informations à donner sur une expérience de puces à ADN. MAGE-OM et MAGE-ML en sont la représentation normalisée, respectivement comme modèle à objets et comme format d'échange.

8.4. Collaborations régionales

Nous avons de multiples collaborations avec les laboratoires de biologie de la région. Ces collaborations sont détaillées dans la section relative aux résultats nouveaux. Parmi les plus avancées, citons

- Inra Rennes - Laboratoire de Génétique Animale : analyse de la régulation de gènes impliqués dans la lipogénèse (M. Le Borgne, J. Nicolas, A. Siegel).
- Inra Rennes - Scribe : analyse de gènes impliqués dans la régulation du stress (M. Le Borgne, J. Nicolas).
- Inra Rennes - Technologie Laitière - Microbiologie : Génomique comparative dans l'étude de la plasticité du génome de Staphylococcus Aureus (R. Andonov, D. Lavenier).
- Inserm Rennes : étude du transcriptome hépatique avec l'unité de recherches hépatologiques (C. Guillouzo, U522) et étude de la famille des défensines avec le laboratoire d'étude de la reproduction chez le mâle, comparaison intensive de séquences (B. Jegou, U435, groupe Germ) (D. Lavenier, J. Nicolas).
- UMR 6026 (Equipe récepteurs et canaux membranaires) : étude de la structure des protéines MIP (C. Belleannée, J. Nicolas, F. Coste, D. Lavenier).
- UMR-CNRS 6061 - Génétique et Développement : Analyse statistique des SNPs (I.-C. Lerman).

8.5. Collaborations nationales

Le projet Symbiose est impliqué dans les programmes nationaux suivants :

- Programme Agenae (E. Retout, Y. Bastide, J. Nicolas) de l'Inra, dirigé par C. Chevalet, avec comme responsables rennais F. Legac pour le transcriptome de la truite et M. Douaire pour celui du poulet. J. Nicolas est par ailleurs membre du conseil scientifique du département BIA de l'Inra (Biométrie et Intelligence Artificielle), réparti sur 6 unités de recherche, principalement en région parisienne et à Toulouse.
- Action IMPG (Informatique, Mathématiques et Physique pour les génomes), dirigé par O. Gascuel (F. Coste, J. Nicolas).
- Réseau Thématique Pluridisciplinaire du département Stic du CNRS *4.1 Bioinformatique : de la séquence génomique à la fonction biologique* (F. Coste, J. Nicolas). J. Nicolas est co-responsable avec T. Lecrocq du thème *Découverte et recherche de motifs* menée au sein de l'action Albio qui correspond au démarrage de ce réseau. Nous démarrons une participation à l'AS *apprentissage et biologie* (Responsables : Jean-Daniel Zucker (Lim&Bio), François Denis (Lif)) et qui se rattache à la fois au RTP 4.1 et au 1.2 (découvrir et résumer).
- Animation scientifique du PCIO (Pole de calcul intensif de l'ouest) (H. Leroy).
- Projet Mathstic 2002 : *Dynamique des réseaux de régulation génique*, animé par A. Siegel et E. Pecou (M. Leborgne, A. Siegel).
- Projet Gafo-Qualité de l'Action Spécifique CNRS Gafo-Données (I.-C. Lerman).

8.6. Collaborations internationales

- Université de Sofia. Suite à la visite de N. Yanev, un travail en collaboration a démarré sur la recherche d'automates minimaux déterministes, sur la reconnaissance du repliement des protéines et sur la plasticité du génome de Staphylococcus Aureus. Les personnes impliquées sont J. Nicolas, F. Coste, R. Andonov, D. Fredouille et D. Lavenier.
- Université de Potsdam (apprentissage dans les réseaux métaboliques). Une thèse en cotutelle a démarré entre nos deux établissements (A. Floëter). Les données proviennent de l'institut Max Planck.
- Université de Genève et SIB (découverte de motifs). Une thèse en cotutelle a démarré fin 2001 (Y. Mescam) avec le Swiss Institute of Bioinformatics (équipe PIG) portant sur la découverte de motifs dans les protéines. Le SIB est un acteur international majeur en protéomique.

- Université de Porto et Lisbonne (analyse de données).
- Université de Liverpool(traitement de données d'expression). Le contrat européen StressGenes nous fournit une première opportunité de développement de relations, concrétisé par une visite de Y. Bastide en octobre pour une semaine.

8.7. Accueils de chercheurs étrangers

- Visite André Floëter (Université de Potsdam) une semaine en mars et novembre.
- Deux visites de quatre jours de Robin Gras et David Hernandez (Université de Genève).
- Visite de Nicolas Yanev, professeur à l'université de Sofia (Bulgarie), pendant 4 mois.

9. Diffusion des résultats

9.1. La conférence JOBIM 2002

Le projet Symbiose a organisé la troisième édition de la conférence JOBIM (Journées Ouvertes Biologie Informatique Mathématiques) à St Malo du 10 au 12 juin 2002 : le comité de programme a été présidé par J. Nicolas et C. Thermes (CGM CNRS) alors que le comité d'organisation a impliqué intensivement la plupart des membres du projet coordonnés par F. Coste et E. Lebret.

Soutenues par l'action ministérielle IMPG (Informatique, Mathématique et Physique pour la Génomique), ces journées constituent le rassemblement annuel de la communauté francophone en bioinformatique. Les thèmes de JOBIM recouvrent tous les domaines de recherche liés à l'analyse et l'exploitation des données génomiques et post-génomiques. Sont inclus parmi ces thèmes : l'analyse des génomes, la modélisation des molécules biologiques, l'analyse des interactions macromoléculaires et l'étude de l'évolution des espèces sur la base de leurs génomes.

Cette troisième édition a confirmé l'importance majeure de ces rencontres dans la communauté : environ 500 personnes ont assisté aux journées, 148 soumissions ont été reçues qui ont permis de proposer 36 communications longues, 35 présentations courtes et 82 posters [12]. Cette reconnaissance dépasse à présent le cadre francophone comme peuvent en témoigner le choix de JOBIM pour accueillir dans sa prochaine édition la conférence ECCB (European Conference on Computational Biology) aussi bien que la provenance de certaines soumissions et la qualité des scientifiques étrangers ayant accepté de donner une conférence invitée :

- *M. Afshar* (UK - RiboTargets), responsable de la conception de médicaments au sein de RiboTarget ;
- *A. Bairoch* (Swiss Institute of Bioinformatics), créateur bien connu de la banque de protéines Swiss-Prot ;
- *D. Gilbert* (UK - School of Informatics), directeur du laboratoire de Bioinformatique de la City University à Londres et *visiting research fellow* à l'EBI, laboratoire européen de bioinformatique, spécialiste d'analyse et découverte de motifs dans les bases de données et en charge du projet TOPs d'analyse de topologies de protéines ;
- *P. Karp* (US - Stanford Research Institute), directeur du groupe de Bioinformatique du laboratoire d'Intelligence Artificielle du SRI, développeur de la base de données EcoCyc ;
- *S. Miyano* (Japon - Human Genome Center), professeur à l'université de Tokyo et directeur du laboratoire *DNA Information Analysis*, concepteur du système d'assistance à la découverte *Hypothesis Creator*.

La qualité des intervenants, le nombre de soumissions et le nombre de participants à cette édition ont ainsi permis de présenter un programme scientifique de qualité tout en préservant, notamment au cours des communications courtes et des séances posters, les échanges d'idées ainsi que les contacts informels essentiels à ce type de rencontre.

9.2. Animation de la communauté scientifique

9.2.1. Animations de revues

- *La Revue de Modulad* (I.-C. Lerman et B. Tallur, comité de lecture).
- *Mathématiques, Informatique & Sciences Humaines*, éditée par le centre d'Analyse et de Mathématiques Sociales (I.-C. Lerman, membre du comité de rédaction).
- *RO-Operations Research* (I.-C. Lerman, éditeur associé).
- *Traitement du Signal* (D. Lavenier, comité éditorial).

9.2.2. Organisation de conférences

- CAp'02 (Conférence d'Apprentissage Francophone), Orléans, 17 au 19 juin 2002 (F. Coste, comité de programme).
- Journées Nationales EGC 2002, Extraction et Gestion de Connaissances, Montpellier, 21-23 janvier 2002 (I.-C. Lerman, comité de programme).
- EGC-2003 (Extraction et Gestion de Connaissances), janvier 2003 (I.-C. Lerman, comité de programme).
- ERSA : International Conference on Engineering of Reconfigurable Systems and Algorithms (D. Lavenier, comités de pilotage et de programme).
- ICGI : International Colloquium Grammatical Inference (F. Coste, comité de pilotage).
- FPL : International Conference on Field Programmable Logic and Applications (D. Lavenier, comité de programme).
- SFC-2002 (Rencontres de la Société Francophone de Classification), septembre 2002, Toulouse (I.-C. Lerman, comité de programme).
- SympA : Symposium en Architectures de Machines (D. Lavenier, comités de pilotage et de programme).

Notons que I.-C. Lerman est membre du conseil d'administration de la société francophone de classification et que F. Coste participe à l'animation de Gowachin, serveur de jeux de test pour l'évaluation de programmes d'inférence grammaticale (http://www.irisa.fr/Gowachin/).

9.3. Enseignements universitaires

Nous participons de façon active à l'enseignement de la bioinformatique. En particulier, M. Le Borgne est chargé de mission auprès de l'Ifsic pour étudier les besoins de formation dans ce domaine en relation avec la partie sciences de la vie de l'Université de Rennes 1. De même, D. Lavenier est co-responsable du DEA Génomique et informatique de l'école doctorale Vie-Santé de l'université de Rennes I. L'originalité du DEA réside dans le double recrutement de biologistes et d'informaticiens. Le DEA est donc ouvert au niveau national aux étudiants titulaires d'une maîtrise de biologie ou d'une maîtrise d'informatique. De nombreux membres du projet participent à cet enseignement.

Un master de bioinformatique à finalité recherche sur deux années est actuellement à l'étude.

Outre le service normal des enseignants du projet, il faut noter les participations suivantes :

1. DEA Génomique et Informatique. (F. Coste, D. Lavenier, I.-C. Lerman, J. Nicolas, B. Tallur).
2. DEA IFSIC. *Classification et Apprentissage* (I.-C. Lerman).
3. DESS MITIC. *Apprentissage sur les textes* (F. Coste) ; *Techniques du data mining* (B. Tallur).
4. DESS Mathématiques Appliquées *Méthodes d'analyse des données- analyse factorielle et classification* (B. Tallur).
5. DIIC. *Reconnaissance des Formes Statistiques* (I.-C. Lerman, B. Tallur) ; *Analyse des données* (B. Tallur).
6. INSA Rennes *Option bioinformatique* (D. Lavenier, J. Nicolas).
7. Formation continue Inra et CNRS (F. Coste, D. Fredouille, D. Lavenier, J. Nicolas).

9.4. Participation à des colloques, séminaires, invitations

9.4.1. Colloques

Nous avons assisté aux manifestations suivantes :

- Réunion Analyse et fouille des données en psychologie (organisée par le LORIA Nancy), février 02, Paris ((I.-C. Lerman, conférencier).
- CAp'02 (Conférence d'Apprentissage Francophone), Orléans (F. Coste).
- ICGI (International Colloquim on Grammatical Inference), Amsterdam, septembre 02 (D. Fredouille, F. Coste).
- IFCS (Conference of the International Federation of Classification Societies), Cracow, Poland, juillet 02 (I.-C. Lerman).
- Forum des jeunes mathématiciennes et des jeunes informaticiennes, Paris, mars 02 (A. Siegel, conférencière).
- journées GRID@Inria, Lyon, janvier 02 (M. Mac Wing).
- journées GRID@Inria, Inria Sophia Antipolis, juillet 02 (H. Leroy, conférencier).
- Groupe de travail GafoQualité, Nantes, juin 02 (I.-C. Lerman).
- Ecole Imaging, Modeling, Manipulating transcriptional regulatory networks, Ambleteuse, octobre 02 (A. Siegel).
- Ecole Modelling and simulation of the biological processes in the context of genomics, Autran, mars 02 (M. Le Borgne).
- OCM 2002 (Objets, Composants, Modèles), Nantes, mars 02 (M. Mac Wing).
- Rencontre Substitutions généralisées, pavages et numération, Marseille, mars 02 (A. Siegel, conférencière).
- SPAA (Symposium on Parallel Algorithms and Architectures), Winnipeg, Canada, août 02 (R. Andonov, conférencier)
- SympA'8 (8ème Symposium en Architectures Nouvelles de Machines), Hamamet, Tunisie, avril 02 (S. Guyetant, conférencier, D. Lavenier).

9.4.2. Invitations

Les membres de Symbiose ont effectué les déplacements suivants :

- Colorado State University. Computer Science Department. Préparation d'un stage de 4 mois de J. Pley à CSD. (R. Andonov, une semaine).
- Université de La Laguna, Espagne. Laboratoire DEIOC. Travail sur la parallélisation de l'ARN. (R. Andonov, un mois invité).
- Genève. SIB. Thèse en co-tutelle. (Y. Mescam, février, juillet et novembre).
- Université de Potsdam, Allemagne. Thèse co-tutelle, lectures in bioinformatics (J. Nicolas, 10 jours, juin).

9.4.3. Exposés invités

- Bangladesh et Inde, septembre. Conférences à Dhaka (Bangladesh) et Delhi, Pilani, Bangalore (Inde), *Bioinformatics Research in France* ; Bangalore (Inde), *Speeding up time consumming genomics application* (D. Lavenier).
- Angers, LERIA, février. *Découverte de motifs en génomique et inférence grammaticale* (J. Nicolas).
- Genève, Swiss Institute of Bioinformatics, Proteome Informatics Group, mars. *Présentation du projet Symbiose* (J. Nicolas) ; *Présentation du projet GenoGRID* (R. Andonov).
- Lyon, avril. Réunion des partenaires du projet européen DATAGRID. *Présentation du projet GenoGRID* (R. Andonov).
- Montpellier, RTP CNRS Bioinformatique (groupe ALBIO IMPG). *Problèmes et méthodes en découverte de motifs* (J. Nicolas).

- Paris, LIAFA, janvier. *Inférence grammaticale de langages réguliers* (D. Fredouille).

10. Bibliographie

Bibliographie de référence

[1] R. ANDONOV, S. BALEV, S. RAJOPADHYE, N. YANEV. *Optimal semi-oblique tiling.* in « SPAA'01 : Proceedings of the Thirteenth annual ACM Symposium on Parallel Algorithms and Architectures », ACM Press, pages 153-162, Crete Island, Greece, 2001.

[2] C. BELLEANNÉE, J. NICOLAS, R. VORC'H. *Vers un démonstrateur adaptatif.* éditeurs J. SALLANTIN, J.-J. SZCZECINIARZ., in « Le concept de preuve à la lumière de l'intelligence artificielle », série Nouvelle Encyclopédie Diderot, Presses Universitaires de France, 1999.

[3] F. COSTE, D. FREDOUILLE. *Efficient ambiguity detection in C-NFA, a step toward inference of non deterministic automata.* in « ICGI 2000, Grammatical inference : algorithms and applications », éditeurs A. L. OLIVEIRA., pages 25-38, Lisbonne, 2000.

[4] J. F. P. DA COSTA, I. C. LERMAN. *Arcade : A Prediction Method for Nominal Variables.*. in « Intelligent Data Analysis (IDA) », numéro 4, volume 2, 1998.

[5] C. DELAMARCHE, P. GUERDOUX-JAMET, R. GRAS, J. NICOLAS. *A symbolic-numeric approach to find patterns in genomes : Application to the translation initiation sites of E. coli.* in « Biochimie », volume 81, 1999.

[6] O. GASCUEL, B. BOUCHON-MEUNIER, G. CARAUX, P. GALLINARI, A. GUÉNOCHE, Y. GUERMEUR, Y. LECHEVALLIER, C. MARSALA, L. MICLET, J. NICOLAS, R. NOCK, M. RAMDANI, M. SEBAG, B. TALLUR, G. VENTURINI, P. VITTE. *Twelve numerical, symbolic and hybrid supervised classification methods.* in « Int. J. of Pattern Recognition and Artificial Intelligence », numéro 5, volume 12, 1998, pages 517-572.

[7] P. GUERDOUX-JAMET, D. LAVENIER. *SAMBA : Hardware Accelerator for Biological Sequence Comparison.* in « CABIOS », numéro 13, volume 6, 1997, pages 609-615.

[8] D. LAVENIER, J. PACHERIE. *Parallel Processing for Scanning Genomic Data-Bases.* in « ParCo'97 (International Conference on Parallel Computing) », Bonn, Germany, 1997.

[9] I. C. LERMAN, P. PETER, J. L. RISLER. *Matrices AVL pour la classification et l'alignement de séquences protéiques.* RR, numéro 2466, Inria, 1994, http://www.inria.fr/rrrt/rr-2466.html.

[10] I.-C. LERMAN, F. ROUXEL. *Comparing classification tree structures : A special case of comparing q-ary relations I & II.* in « RAIRO Operations Research », volume 33 & 34, 1999, pages 339-365 & 251-281.

[11] B. TALLUR, J. NICOLAS, A. FROGER, D. THOMAS, C. DELAMARCHE. *Sequence classification of water channels and related proteins in view of functional predictions.* in « Theoretical chemistry accounts », 1998.

Livres et monographies

[12] *JOBIM : Journées Ouvertes Biologie Informatique Mathématiques.* éditeurs J. NICOLAS, C. THERMES., St-Malo, France, 2002.

[13] N. PYTHEAS-FOGG. *Substitutions in Dynamics, Arithmetics and Combinatorics.* Lectures Notes in Mathematics 1794, Springer-Verlag, 2002, Edité par V. Berthé, S. Ferenczi, C. Mauduit et A. Siegel.

Articles et chapitres de livre

[14] R. ANDONOV, S. BALEV, S. RAJOPADHYE, N. YANEV. *Optimal semi-oblique tiling and its application to sequence comparison.* in « Transactions on Parallel and Distributed Systems », 2002, à paraîtreRapport Interne 2001 Irisa 1392.

[15] V. BERTHÉ, A. SIEGEL. *1. Basic notions on substitutions.* in « Substitutions in Dynamics, Arithmetics and Combinatorics, N. Pytheas-Fogg », série Lectures Notes in Mathematics 1794, Springer-Verlag, 2002, pages 1-34.

[16] A. ELAMRANI, L. MARIE, A. AÏNOUCHE, J. NICOLAS, I. COUÉE. *Genome wide distribution and potential regulatory functions of AtATE, a novel miniature inverted-repeat transposable element that is present in the promoter region of one of the Arginine Decarboxylase genes in Arabidopsis Thaliana.* in « Molecular Genetics and Genomics », volume 267, 2002, pages 459-471.

[17] I.-C. LERMAN, P. PETER. *Indice probabiliste de vraisemblance du lien entre objets quelconques : analyse comparative entre deux approches.* in « Revue de Statistique Appliquée », 2002, à paraître.

[18] A. SIEGEL. *7. Spectral theory and geometric representation of substitutions.* in « Substitutions in Dynamics, Arithmetics and Combinatorics, N. Pytheas-Fogg », série Lectures Notes in Mathematics 1794, Springer-Verlag, 2002, pages 199-252.

[19] A. SIEGEL. *Représentation des systèmes dynamiques substitutifs non unimodulaire.* in « Ergodic Theory and Dynamical Systems », 2002, à paraître.

[20] A. SIEGEL. *Répétitions dans les figures géométriques.* in « Quadrature », numéro 34, 2002, pages 40-47.

Communications à des congrès, colloques, etc.

[21] F. ALMEIDA, R. ANDONOV, D. GONZALEZ, L. MORENO, V. POIRRIEZ, C. RODRIGUEZ. *Optimal tiling for the RNA base pairing problem.* in « 14th ACM Symposium on Parallel Algorithms and Architectures (SPAA) », pages 173-182, Winnipeg, Canada, 2002.

[22] J. COSTA, I. LERMAN, H. SILVA. *Linéarisation d'un critère de classification en cas de données numériques et qualitatives nominales.* in « Actes du 8-ème congrès de la Société Francophone de Classification, 2001 », pages 99-106, Pointe-à-Pitre, Guadeloupe, 2002.

[23] S. GUYÉTANT, S. DERRIEN, D. LAVENIER. *Architecture parallèle pour la génomique.* in « SympA'8. 8ème Symposium en Architectures Nouvelles de Machines », pages 361-364, Hamamet, Tunisie, 2002.

[24] D. LAVENIER, H. LEROY, M. HURFIN, R. ANDONOV, L. MOUCHARD, F. GUINAND. *Le projet GénoGRID : une grille expérimentale pour la génomique.* in « JOBIM 2002. Journées Ouvertes Biologie Informatique Mathématiques », pages 27-31, Saint Malo, France, 2002.

[25] I. LERMAN, K. BACHAR. *Agrégations multiples et contraintes de contiguïté dans la classification ascendante hiérarchique utilisant les voisins réciproques et le critère de la vraisemblance des liens.* in « Actes du 8-ème congrès de la Société Francophone de Classification, 2001 », pages 232-237, Pointe-à-Pitre, Guadeloupe, 2002.

[26] I.-C. LERMAN, J. P. DA COSTA, H. SILVA. *Validation of Very Large Data Sets Clustering by Means of a Nonparametric Linear Criterion.* in « Classification, Clustering and Data Analysis. Recent Advances and Applications », Springer-Verlag, éditeurs A. S. K. JAJUGA, H.-H. BOCK., pages 147-157, 2002.

[27] J. PLEY, R. ANDONOV, J.-F. GIBRAT, A. MARIN, V. POIRRIEZ. *Parallélisation d'une méthode de reconnaissance de repliements de protéines (poster).* in « JOBIM 2002. Journées Ouvertes Biologie Informatique Mathématiques », pages 287-288, Saint Malo, France, 2002.

[28] F. RAIMBAULT, D. LAVENIER. *ROOM : des machines reconfigurables orientés objet.* in « SympA'8. 8ème Symposium en Architectures Nouvelles de Machines », pages 346-353, Hamamet, Tunisie, 2002.

[29] A. SIEGEL. *Autour des fractals de Rauzy.* in « Forum des jeunes mathématiciennes et des jeunes informaticiennes », pages 77-80, 2002.

Rapports de recherche et publications internes

[30] C. ALLAND. *Plate-forme d'extraction de motifs dans des ensembles de séquences génomiques.* Rapport d'activité, Irisa, 2002, http://www-interne.irisa.fr/atelier/documentations/ASCII/travauxIA/symbiose/ra.htm.

[31] Y. BASTIDE, J. NICOLAS. *Stressgenes : progress report at 6 months.* Rapport intermédiaire, Commission Européenne, 2002.

[32] I.-C. LERMAN, J. AZÉ. *Indice Probabiliste Discriminant (de vraisemblance du lien) d'une Règle d'Association en cas de "Très Grosses" Données.* rapport technique, CNRS, Action Spécifique STIC "Gafo-Données", 2002.

[33] I.-C. LERMAN, P. PETER. *Indice probabiliste de vraisemblance du lien entre objets quelconques : analyse comparative entre deux approches.* Rapport Interne, numéro 1438, Irisa, 2002.

[34] A. SIEGEL. *Pure discrete spectrum dynamical system and periodic tiling associated with a substitution.* Prépublication, Irisa, 2002, http://www.irisa.fr/symbiose/people/siegel/Pro/publi.htm.

[35] N. YANEV, R. ANDONOV. *The protein threading problem is in P ?.* RR, numéro 4577, Inria, 2002, http://www.inria.fr/rrrt/rr-4577.html, soumis à RECOMB-2003.

Bibliographie générale

[36] P. ARNOUX, S. ITO. *Pisot substitutions and Rauzy fractals.* in « Bull. Belg. Math. Soc. Simon Stevin »,

numéro 2, volume 8, 2001, pages 181-207, Journées Montoises d'Informatique Théorique (Marne-la-Vallée, 2000).

[37] K. BACHAR. *Contributions en analyse factorielle et en classification ascendante hiérarchique sous contrainte de contiguïté. Applications à la segmentation d'images.* thèse de doctorat, Université de Rennes 1, 1994.

[38] A. BRAZMA, P. HINGAMP, J. QUACKENBUSH, G. SHERLOCK, P. SPELLMAN, C. STOECKERT, J. AACH, W. ANSORGE, C. A. BALL, H. C. CAUSTON, T. GAASTERLAND, P. GLENISSON, F. C. HOLSTEGE, I. F. KIM, V. MARKOWITZ, J. C. MATESE, H. PARKINSON, A. ROBINSON, U. SARKANS, S. SCHULZE-KREMER, J. STEWART, R. TAYLOR, J. VILO1, M. VINGRON. *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* in « Nature Genetics », numéro 4, volume 29, décembre, 2001, pages 365-371.

[39] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Efficient discovery of conserved patterns using a pattern graph..* in « Cabios », numéro 13, 1997, pages 509-522.

[40] A. BRAZMA, I. JONASSEN, I. EIDHAMMER, D. GILBERT. *Approaches to the Automatic Discovery of Patterns in Biosequences.* in « Journal of Computational Biology », numéro 2, volume 5, 1998, pages 277-304.

[41] BRYANT. *Graph-based algorithms for boolean function manipulation.* in « IEEE Transactions on Computers », numéro 8, volume C, 1986, pages 677-691.

[42] J. BUHLER, M. TAMPA. *Findind motifs using random projections.* in « Proceedings of RECOMB01 », ACM Press, pages 69-76, Montreal, Canada, 2001.

[43] V. CANTERINI, A. SIEGEL. *Geometric representation of substitutions of Pisot type.* in « Trans. Amer. Math. Soc. », numéro 12, volume 353, 2001, pages 5121-5144.

[44] E. CHOW, T. HUNKAPILLER, J. PETERSON. *Biological Information Signal Processor.* in « ASAP », 1991, pages 144-160.

[45] E. CLARKE, D. LONG, K. MCMILLAN. *A language for compositional specification and verification of finite state harware controllers.* in « Proceeding of the IEEE », numéro 9, volume 79, 1991, pages 1283-1292.

[46] J. COLLADO-VIDES. *A Transformational-Grammar Approach to the Study of The Regulation of Gene Expression.* in « J. Theor. Biol. », numéro 6, volume 13, 1989, pages 403-425.

[47] O. COUDERT, J. MADRE. *A unified framework for the formal verification of sequential circuits.* in « Conference on Computer-aided Design », pages 126-130, 1990.

[48] H. DE JONG. *Modeling and Simulation of genetic Regulatory Systems : a Literature Review.* in « Journal of Computational Biology », volume 9 (1), 2002, pages 69-105.

[49] H. DE JONG, M. PAGE. *Qualitative simulation of large and complex genetic regulatory systems.* in « Proceeding of the 14th European Conference on Artificial Intelligence, ECAI 2000 », IOS Press, éditeurs W. HORN., pages 141-145, Amsterdam, 2000.

[50] S. DONG, D. SEARLS. *Gene structure prediction by linguistic methods.* in « Genomics », volume 23, 1994, pages 540-551.

[51] N. FRIEDMAN, M. LINIAL, I. NACHMAN, D. PE'ER. *Using Bayesian Ntworks to analyse expression data.* in « Proceedings of the Fouth Annual International Conference on Computational Molecular Biology, RECOMB2000 », ACM Press, New-York, N.Y, 2000.

[52] A. FROGER, B. TALLUR, D. THOMAS, C. DELAMARCHE. *Prediction of functional residues in water channels and related proteins.* in « Protein Science », volume 7, 1998, pages 1458-1468.

[53] E. GLEMET, J. CODANI. *LASSAP : a LArge Scale Sequence compArison Package,.* in « Cabios », numéro 2, volume 13, 1997, pages 137-143.

[54] R. GRAS. *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques.* Thèse d'état, Université de Rennes 1, 1979.

[55] R. GRAS. *Un outil interactif de recherche de motifs dans les grandes séquences génétiques fondé sur l'arbre des suffixes.* thèse de doctorat, Université de Rennes I, 1997.

[56] T. HEAD. *Formal language theory and DNA : an analysis of the generative capacity of specific recombinant behaviours.* in « Bull. Math. Biology », volume 49, 1987, pages 737-759.

[57] J. HENIKOFF, S. HENIKOFF. *BLOCKs database and its applications.* in « Methods Enzymol. », volume 266, 1996, pages 88-105.

[58] Z. HUANG. *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.* in « Data Mining and Knowledge Discovery, Kluwer », volume 2, 1998.

[59] J. HUDAK, M. MCCLURE. *A comparative analysis of computational motif-detection methods.* in « Pacific Symposium of Biocomputing PSB 1999 », pages 138-139, 1999, http://www-smi.stanford.edu/projects/helix/psb99.

[60] N. JAMSHIDI, S. JEREMY, J. EDWARD, T. FAHLAND, G. CHURCH, B. PALSSON. *Dynamic simultion of the human red blood cell metabolic network..* in « Bioinformatics », volume 17, 2001, pages 286-287.

[61] I. JONASSEN, J. COLLINS, D. HIGGINS. *Finding flexible patterns in unaligned protein sequences.* in « Protein Science », numéro 8, volume 4, 1995, pages 1587-1595.

[62] I. JONASSEN. *Efficient discovery of conserved patterns using a pattern graph.* in « Cabios », volume 13, 1997, pages 509-522.

[63] M. KANEHISA, S. GOTO. *KEGG : Kyoto Encyclopedia of Genes and Genomes..* in « Nucleic Acids Research », numéro 1, volume 28, 2000, pages 27-30.

[64] L. KARI, G. PAUN, G. ROZENBERG, A. SALOMAA, S. YU. *DNA computing, Sticker systems and universality.* in « Acta Informatica », volume 35, 1998, pages 401-420.

[65] P. KARP, M. RILEY, S. PALEY, A. PELLEGRINI-TOOLE, M. KRUMMENACKER. *EcoCyc : Encyclopedia of Escherichia coli gens and metabolism.* in « Nucleic Acids Research », numéro 1, volume 27, 1999, pages 55-58.

[66] S. KAUFFMAN. *The large scale structure and dynamics of gene control circuits : an ensemble approach.* in « Journal of Theorical biology », volume 44, 1974, pages 167.

[67] V. KEICH, A. PEVZNER. *Findind motifs in the twilight zone.* in « Proceedings of RECOMB02 », ACM Press, pages 195-203, Washington, USA, 2002.

[68] F. KOLPAKOV, E. ANANKO, G. KOLESOV, N. KOLCHANOV.. *GeneNet : A gene network database and its automated visualisation.* in « Bioinformatics », numéro 6, volume 14, 1999.

[69] A. LANDRAUD, J. AVRIL, P. CHRETIENNE. *An algorithm for finding a common structure shared by a family of strings.* in « IEEE », numéro 8, volume 11, 1989, pages 890-895.

[70] C. E. LAWRENCE, S. F. ALTSCHUL, M. S. BOGUSKI, J. S. LIU, A. F. NEUWALD, J. C. WOOTTON. *Detecting subtle sequence signals : a Gibbs sampling strategy for multiple alignment..* in « Science », volume 262, 1993, pages 208-214.

[71] T. LENGAUER. *Bioinformatics. From genoms to Drugs.* Wiley-VCH, 2002.

[72] L. MARSAN, M.-F. SAGOT. *Algorithms for extracting structured motifs using a suffix-tree with application to promoter and regulatory site consensus identification.* in « J. of Comput. Biol. », numéro 7, 2001, pages 345-360.

[73] L. MENDOZA, D. THIEFFRY, E. ALVAREZ-BUYLLA. *Genetic control of flower morphogenesis in Arabidopsis thaliana : a logical analysis.* in « Bioinformatics », numéro 7/8, volume 15, 1999, pages 593-606.

[74] S. NEEDLEMAN, C. WUNSCH. *A general method applicable to the search of similarities in the amino acid sequences of two protein,.* in « J. Mol. Biol. », volume 48, 1970, pages 443-453.

[75] G. PAUN, G. ROZENBERG, A. SALOMAA. *DNA Computing. New Computing Paradigms.* Springer-Verlag, 1998.

[76] A. PEVZNER, S.-H. SZE. *Combinatorial approaches to finding subtle signals in DNA sequence.* in « Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB) », pages 269-278, 2000.

[77] M. QUEFFÉLEC. *Substitution dynamical systems-spectral analysis.* Lecture Notes in Mathematics, 1294. Springer-Verlag, Berlin, 1987.

[78] G. RAUZY. *Nombres algébriques et substitutions.* in « Bull. Soc. Math. France », numéro 2, volume 110, 1982, pages 147-178.

[79] A. REGEV, W. SILVERMAN, E. SHAPIRO. *Representation and simulation of biochemical processes using the pi-calculus process algebra.* in « Pacific Symposium of Biocomputing 2001 », 2001.

[80] M.-F. SAGOT, A. VIARI. *A Double Combinatorial Approach to Discovering Patterns in Biological Sequences.* in « Proceedings of the7th Annual Symposium on Combinatorial Pattern Matching », série 1075, Springer-Verlag, Berlin, éditeurs D. S. HIRSCHBERG, E. W. MYERS., pages 186-208, Laguna Beach, CA, 1996.

[81] Y. SAKAKIBARA. *Recent advances of grammatical inference.* in « Theoretical Computer Science », volume 185, 1997, pages 15-45.

[82] H. SALGADO, A. SANTOS, U. GARZA-RAMOS, J. VAN HELDEN, E. DIAZ, J. COLLADO-VIDES. *RegulonDB(version 2.0) : A database on transcriptional regulation in Escherichia coli..* in « Nucleic Acids Research », numéro 1, volume 27, 2000, pages 59-60.

[83] D. B. SEARLS. *String Variable Grammar : A Logic Grammar Formalism for the Biological Language of DNA.* in « Journal of Logic Programming », numéro 1/2, volume 24, 1995, pages 73-102.

[84] D. SEARLS. *Formal language theory and biological macromolecules.* in « Theoretical Computer Science », volume 47, 1999, pages 117-140.

[85] E. SNOUSSI. *Necessary conditions for multistationnarity and stable periodicity.* in « J. Biol. Syst. », volume 6, 1998, pages 1-23.

[86] R. STADEN. *Methods for discovering novel motifs in nucleic acid sequences.* in « Comput. Applic. Biosci. », volume 5, 1989, pages 89-96.

[87] D. STATES, W. GISH, S. ALTSCHUL. *Basic local alignment search tool,.* in « J. Mol. Biol. », volume 215, 1990, pages 403-410.

[88] B. TALLUR, J. NICOLAS. *A method for classifying unaligned biological séquences.* in « IFCS-96 (Data Science, Classification and Related Methods) », Springer Verlag, Tokyo, 1997.

[89] R. THOMAS, D. THIEFFRY, M. KAUFFMAN. *Dynamical behaviour of biological regulatory networks-I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state.* in « Bull. Math. Biol. », volume 57, 1995, pages 247-276.

[90] R. THOMAS. *Regulatory networks seen as asynchronous automata : a logical description.* in « J. Theor. Biol. », volume 153, 1991, pages 1-23.

[91] C. WHITE, R. SINGH, P. REINTJES, J. LAMPE, B. ERICKSON, W. DETTLOFF, V. CHI, S. ALTSCHUL. *BioSCAN : A VLSI-Based System for Biosequence Analysis,.* in « IEEE Int. Conf on Computer Design : VLSI in Computer and Processors », pages 504-509, 1991.

[92] T. YOKOMORI, S. KOBAYASHI. *DNA Evolutionary Linguistics and RNA Structure Modeling : A Computational Approach.* in « Proc.of 1st International IEEE Symposium on Intelligence in Neural and Biological Systems », pages 38-45, 1995.